Analyse richtungsabhängiger $\mathcal{H}^2$-Matrizen

Dissertation
zur Erlangung des Doktorgrades
der Mathematisch-Naturwissenschaftlichen Fakultät
der Christian-Albrechts-Universität zu Kiel

vorgelegt von
Christina Börst

Kiel, 2021

Erstgutachter:	Prof. Dr. Steffen Börm
Zweitgutachter:	Prof. Dr. Jens Markus Melenk
Tag der mündlichen Prüfung:	18.10.2021

Zusammenfassung

Ziel dieser Arbeit ist es, einen Überblick über die Grundlagen der $\mathcal{RH}^2$-Matrizen zu bieten und die Matrizen selbst sowie einige Algorithmen zu analysieren. $\mathcal{RH}^2$-Matrizen wurden entwickelt, da herkömmliche $\mathcal{H}^2$-Matrizen beim Anwenden auf die hochfrequente Helmholtz-Gleichung hohe Ränge benötigen und damit auf die Grenzen ihrer Effizienz stoßen.
Die Grundidee ist es, die Kernfunktion durch eine ebene Welle entlang einer vorgegebenen Richtung zu dividieren und auf diese Weise eine modifizierte Kernfunktion zu erhalten, welche innerhalb eines Kegels glatt ist. Innerhalb des Kegels sind somit keine hohen Ränge für die Approximation notwendig. Zum Abdecken des gesamten Gebiets werden mehrere Kegel und entsprechend mehrere Richtungen für die ebenen Wellen benötigt. Die Modifikation der Kernfunktion macht es erforderlich, das Verhalten der Kernfunktion erneut zu analysieren, um Fehlerabschätzungen für die Approximation zu gewinnen. Ebenso machen es die auftretenden Richtungen notwendig, neue Aufwandsabschätzungen zu erstellen und Algorithmen zu erarbeiten.

Erste Analysen des Approximationsfehlers und des Aufwands für den Fall des Einfachschichtoperators wurden von Melenk und Börm vorgenommen. Die vorliegende Arbeit baut auf diesen Ergebnissen auf und versucht, einige Aspekte wie die Schwachbesetztheit der $\mathcal{RH}^2$-Matrizen aus einem anderen Blickwinkel zu betrachten. Im Zuge der Arbeit wird die Fehleranalyse auf Teilmatrizen für den Einfachschichtoperator und erste Analysen für den Doppelschichtoperator erweitert. Die von der verwendeten Wellenzahl κ unabhängig exponentielle Konvergenz des Interpolationsfehlers bleibt auch im Fall des Doppelschichtoperators erhalten.
Des Weiteren werden Konzepte zur Minimierung des Rangs sowie Vergröberung der zugrundeliegenden Baumstruktur übertragen und entwickelt. Zur Minimierung der Ränge wird ein Algorithmus zur Orthogonalisierung von Clusterbasen für $\mathcal{RH}^2$-Matrizen ausgearbeitet und analysiert, welcher zudem in der Programmbibliothek *H2Lib* (die Programmbibliothek der Arbeitsgruppe *scientific computing*) implementiert wurde. Aufbauend auf der Orthogonalisierung entsteht ein Algorithmus zur Rekompression von $\mathcal{RH}^2$-Matrizen, der ebenfalls analysiert wird und in der Programmbibliothek implementiert wurde. Die Rekompression führt zu einer deutlichen Reduktion der Ränge und damit zu erheblich geringeren Speicheranforderungen, teilweise auf ein Zehntel, der Aufwand beträgt dabei $\mathcal{O}(k^2(n+\kappa^2 k\log_2(n)))$. Um eine weitere Reduktion des Aufwands zu erreichen, wird eine Variante der Vergröberung des Blockbaums entwickelt und analysiert. Auch dieser Algorithmus wurde in die Programmbibliothek integriert.

Abstract

The goal of this dissertation is to analyze $\mathcal{RH}^2$-matrices in greater depth as well as give a general view of their properties and introduce several algorithms. $\mathcal{RH}^2$-matrices were invented to overcome the obstacle of large ranks in the case of the high-frequency Helmholtz equation treated with hierarchical matrices. The basic idea is to divide the kernel function by a plane wave to obtain a function which is smooth inside a cone, since the approximation of a smooth function does not lead to large ranks. However more than one cone is necessary and the kernel function has been modified, therefore the main work is to show that the approximation reaches the desired accuracy, the storage cost is bounded, and efficient algorithms for setting up a matrix and standard matrix arithmetic exist.

A first answer for the question if these conditions still hold was given by Melenk and Börm, but only for the single-layer case.
This work adds a result for the double-layer case where it turns out that the approximation error still shows an exponential convergence rate independent on the wave number. Further the ideas of orthogonal clusterbasis and recompression are transferred and examined for more efficient approximations. Both approaches are implemented in the H2Lib and leads to reduced storage costs. The complexity of both algorithm is bounded by $\mathcal{O}(k^2(n + \kappa^2 k \log_2(n)))$. To overcome the obstacle of large block trees an algorithm for coarsening is created, implemented and tested to analyze its behavior.
In order to verify given statements about the effort of introduced algorithms and approximation errors some experiments are included.

Veröffentlichungen

Der Algorithmus sowie eine Aufwandsabschätzung für die Rekompression wurden vorab im folgenden Artikel in Zusammenarbeit mit Prof. Dr. Börm veröffentlicht. Alle neuen Erkenntnisse dieses Artikels sind in dieser Arbeit enthalten, jedoch detaillierter formuliert.

[10] BÖRM, S. ; BÖRST, C. : Hybrid matrix compression for high-frequency problems. In: *SIAM Journal on Matrix Analysis and Applications* 41 (2020), Nr. 4, S. 1704–1725. DOI 10.1137/19M124280X

Danksagung

Die vorliegende Dissertation wäre ohne die Unterstützung zahlreicher Personen in dieser Form nicht möglich gewesen. Für die vielfältige Hilfe möchte ich mich auf diesem Wege herzlich bedanken auch bei denen, die ich aus Platzgründen hier nicht namentlich aufführen kann.

Mein besonderer Dank gilt zunächst meinem Doktorvater Prof. Dr. Steffen Börm, der mich erst auf die Idee zu promovieren gebracht hat und immer ein offenes Ohr bei Fragen und Problemen hatte.
Danken möchte ich auch seiner Arbeitsgruppe, die mich aufgenommen und über die Zeit hinweg mit hilfreichem Feedback, Denkanstößen oder in schwierigen Phasen auch einfach mal mit ein bisschen *Ablenkung* versorgt hat.
Gesondert möchte ich mich auch bei meinem ehemaligen Kollegen Daniel Hans bedanken, der zu Beginn meiner Arbeit unzählige Stunden mit mir über verschiedene Ideen und Ansätze diskutiert hat, sowie Jonas Lorenzen, der mich intensiv bei der Suche nach hartnäckigen Fehlern und mit alternativen Herangehensweisen unterstützt hat.
Weiterhin danke ich meinen Freundinnen Maraike Schmidt und Lia Rasim, die in ihrer Freizeit das Korrekturlesen meiner Arbeit übernommen und dabei einige langlebige Verschreiber entdeckt haben, ebenso wie meiner Freundin Sarah David, die mir an Tiefpunkten und nach Rückschlägen viele Liter Kaffee und mehr als ein offenes Ohr geschenkt hat.
Für den unerschütterlichen Glauben meiner Mutter Regina Börst in mich, möchte ich ihr herzlich Danken, sie ist eine schier grenzenlose Motivationsquelle.
Schließlich gilt mein ganz besonderer Dank meinem (mittlerweile) Ehemann Hendrick Börst, der mir jederzeit den Rücken frei gehalten und mich auf diese Weise mit der nötigen Zuversicht und Kraft versorgt hat, um dieses Projekt zu beenden.

Inhaltsverzeichnis

Vorwort

Eine Fragestellung, welcher sich Wissenschaftler in den letzten Jahren immer wieder angenommen haben, ist die Konstruktion von Lösungsverfahren für die Helmholtz-Gleichung im hochfrequenten Bereich. Dieser Bereich ist bei klassischen Ansätzen der Paneelclusterungstechniken [31], Multipolentwicklungen [28, 44] oder hierarchischen Matrizen [30, 29] problematisch. Aufgrund der auftretenden hohen lokalen Ränge beziehungsweise der sehr hohen benötigten Auflösung steigt der Aufwand der Methoden so weit an, dass diese nicht mehr als effizient betrachtet werden können. Dies führt dazu, dass hohe Rechenzeiten von Nöten sind und bei manch einer Problemstellung auch moderne Computer an ihre Grenzen stoßen. Um die Rechenzeit drastisch zu reduzieren und für sonst nicht berechenbare Probleme Lösungen mit adäquater Genauigkeit bestimmen zu können, sind neue, angepasste und ausgefeiltere Ansätze von Nöten.

Aktuelle Verfahren lassen sich grob in drei unterschiedliche Typen einteilen, die hier kurz vorgestellt werden sollen.
Die unter dem Sammelbegriff *(Multilevel) Fast Multipole Methods ((M)FMM)* [16] [1] zusammengefassten Methoden haben ihren Ursprung in einer für die Laplace-Gleichung von Rokhlin [44] vorgestellten und dann von Rokhlin sowie Greengard für Partikelsimulationen [28] angepassten Multipolentwicklung. In den folgenden Jahren erfolgten weitere Anpassungen für den Fall der hochfrequenten Helmholtz-Gleichung [45]. FMM können den Berechnungsaufwand im Laplace-Fall für n-Teilchen-Systeme von $\mathcal{O}(n^2)$ auf $\mathcal{O}(n)$ senken, im Fall der hochfrequenten Helmholtz-Gleichung steigt der Aufwand von $\mathcal{O}(n)$ multiplikativ um zusätzliche logarithmische Terme.
Jedoch gestalten sich sowohl die Implementierung als auch das Verallgemeinern der Fehleranalyse als sehr aufwändig, da sie stark an der auftretenden Kernfunktion orientiert sind.

Bei butterfly-Algorithmen [40] wird eine schnelle Matrix-Vektor-Multiplikation dadurch erreicht, dass die Matrix hierarchisch in Teilmatrizen unterteilt und die Multiplikation der Teilblöcke ähnlich wie die schnelle Fourier-Transformation gehandhabt wird. Das Bearbeiten von Teilmatrizen erfolgt nach bestimmten Regeln und kann graphisch mit dem namensgebenden Schmetterlingsgraphen dargestellt werden. Diese Methode erreicht eine Komplexität von $\mathcal{O}(n \log^2(n))$. Die entstehende Approximation ist strukturell simpel und eng mit der der $\mathcal{H}$-Matrix verwandt. Sowohl butterfly-Algorithmen als auch $\mathcal{H}$-Matrizen erreichen jedoch nicht die Kompressionsraten, die bei FMM möglich sind.

Eine weitere Möglichkeit stellen die *direkten Methoden* [12] [23] [2] dar, welche die Eigenschaft der Kernfunktion, nämlich als Produkt mit einer ebenen Welle darstellbar zu sein, unmittelbar nutzen. Dies führt zu einer degenerierten Darstellung der Kernfunktion und infolgedessen zu schnellen Summations-Ansätzen. Jedoch sind die Verfahren maßgeschneidert auf die Problemstellung und damit nicht auf andere Probleme übertragbar.

Die richtungsabhängigen $\mathcal{H}^2$-Matrizen (kurz $\mathcal{RH}^2$-Matrizen), die Gegenstand dieser Arbeit sind, zählen zu den direkten Methoden und sind eine Variation der $\mathcal{H}^2$-Matrizen.

Ziel dieser Arbeit ist eine Vertiefung und Weiterführung der bisherigen Analyse der $\mathcal{RH}^2$-Matrizen [6], [3], [10].
Bevor jedoch eine $\mathcal{RH}^2$-Matrix erstellt werden kann, gilt es, die zugrundeliegende Problemstellung die Helmholtz-Gleichung im Bezug auf Existenz und Eindeutigkeit einer Lösung zu untersuchen und die Problemstellung in eine diskrete Formulierung zu überführen. Das erste Kapitel widmet sich diesen Themen und soll dabei einen kurzen Überblick geben, was es auf der Seite der Problemstellung zu bedenken gibt.
Das daran anschließende Kapitel beschäftigt sich mit den Grundlagen, die für die hier gewählt Approximationsmethode der $\mathcal{RH}^2$-Matrizen notwendig sind. Das heißt, in diesem Kapitel werden die Tensorinterpolation vorgestellt, Bedingungen für die Approximation formuliert und richtungsabhängige Cluster- sowie Blockbäume eingeführt.
In dem dritten Kapitel ist dann alles beisammen, um die $\mathcal{RH}^2$-Matrizen zu definieren, den Speicheraufwand zu analysieren und den wichtigsten Algorithmus, die Matrix-Vektor-Multiplikation, vorzustellen und ebenfalls zu analysieren.
Im darauf folgenden Kapitel werden die auftretenden Interpolationsfehler für den Einfach- und Doppelschichtoperator untersucht und erste Aussagen zu globalen Approximationsfehlern getroffen.
Da die Interpolation zwar einen schnellen, aber nicht unbedingt einen effizienten Zugang zu einer Approximation liefert, werden im anschließenden Kapitel algebraische Optimierungsansätze für den Speicheraufwand vorgestellt.
Während sich die Optimierungsansätze im fünften Kapitel auf das Reduzieren des Rangs durch Berechnung verbesserter Clusterbasen beschränken, werden im letzten Kapitel Verkleinerungen des auftretenden richtungsabhängigen Blockbaums thematisiert.

1 Die Aufgabenstellung

Bevor die $\mathcal{RH}^2$-Matrizen angewendet werden können, gilt es, die Helmholtz-Gleichung zu diskretisieren. Dazu soll zunächst ein kurzer Blick auf die Helmholtz-Gleichung und auf einige mögliche Problemstellungen gerichtet werden. Anschließend folgt eine kurze Einführung in die Randelementmethoden.

1.1 Die Helmholtz-Gleichung

Die Helmholtz-Gleichung[i] lässt sich aus der Wellengleichung herleiten, bei der es sich um eine hyperbolische partielle Differentialgleichung handelt, die sowohl zeit- als auch ortsabhängig ist. Die Helmholtz-Gleichung kann als zeitunabhängige (engl. *steady state*) Variante der Wellengleichung betrachtet werden. Für $d \in \mathbb{N}$, $t \in \mathbb{R}_{\geq 0}$, $x \in \mathbb{R}^d$, $c \in \mathbb{R}_{>0}$ und eine reell- oder komplexwertige Funktion $w : \mathbb{R}_{\geq 0} \times \mathbb{R}^d \to \mathbb{C}$ ist die Wellengleichung durch

$$\frac{1}{c^2} \frac{\partial^2}{\partial t^2} w(t,x) - \sum_{i=1}^{d} \frac{\partial^2}{\partial x_i^2} w(t,x) = 0$$

gegeben. Physikalisch ist die Konstante c als Geschwindigkeit zu interpretieren, die je nach Wellenart und Medium, in dem sich die Welle ausbreitet, zu wählen ist. Wird die Gleichung, zum Beispiel zur Beschreibung einer akustischen Welle herangezogen, entspricht c der jeweiligen Schallgeschwindigkeit in dem betrachteten Medium.
Die Helmholtz-Gleichung ist eine elliptische partielle Differentialgleichung, die zeitharmonische Wellen beschreibt, so zum Beispiel akustische Wellen, die sich bei kontinuierlichen Quellen bilden. Im Zuge dieser Arbeit soll die homogene Helmholtz-Gleichung für eine komplexwertige Funktion $u : \mathbb{R}^3 \to \mathbb{C}$ als Problemstellung dienen. Die Gleichung wird für eine reelle *Wellenzahl* $\kappa \in \mathbb{R}_{\geq 0}$ in der Form

$$0 = -\Delta u(x) - \kappa^2 u(x), \tag{1.1.1}$$

wobei Δ den Laplace-Operator[ii], also $\Delta u = \nabla \cdot \nabla u$ bezeichne, notiert.
Bei der hier betrachteten Wellenzahl handelt es sich physikalisch um die Kreiswellenzahl,

[i] Benannt nach dem deutschen Arzt und Physiker Hermann von Helmholtz, der unter anderem auf dem Gebiet der elektromagnetischen Wellen forschte [25].

[ii] Benannt nach dem französischen Physiker und Mathematiker Pierre-Simon Laplace, welcher sich mit Wahrscheinlichkeitsrechnung und Differentialgleichungen befasste.

welche mit der Kreiszahl π und der Wellenlänge λ im folgenden Verhältnis steht [17, S. 90]

$$\kappa = \frac{2\pi}{\lambda}. \tag{1.1.2}$$

Damit beschreibt die Wellenzahl κ die Anzahl der Schwingungen auf einem Abschnitt der Länge 2π.
Die Helmholtz-Gleichung lässt sich mit Hilfe eines Separationsansatzes mit der Annahme, dass zwei Funktionen $v(t)$ und $u(x)$ derart existieren, dass $w(t,x) = v(t)u(x)$ gilt, aus der Wellengleichung gewinnen [13, S. 175]. Sei Δ_x der Laplace-Operator für x, dann führt der Separationsansatz zu

$$\frac{1}{c^2}\frac{\partial^2}{\partial t^2}v(t)u(x) - \Delta_x v(t)u(x) = 0.$$

Wird zusätzlich $v(t)u(x) \neq 0$ für alle $t \in \mathbb{R}_{\geq 0}$ und alle $x \in \mathbb{R}^d$ gefordert, kann die Gleichung umsortiert werden

$$\frac{1}{c^2}\frac{\frac{\partial^2 v(t)}{\partial t^2}}{v(t)} = \frac{\Delta_x u(x)}{u(x)}.$$

Die linke Seite ist ortsunabhängig, während die rechte unabhängig von der Zeit ist. Somit existiert eine Konstante K, abhängig davon ob w reell oder komplex ist, aus $\mathbb{R}$ oder $\mathbb{C}$ mit

$$\frac{1}{c^2}\frac{\frac{\partial^2 v(t)}{\partial t^2}}{v(t)} = -K = \frac{\Delta_x u(x)}{u(x)}.$$

Aufsplitten in zwei Gleichungen erzeugt eine zeit- und eine ortsabhängige Differentialgleichung

$$\frac{1}{c^2}\frac{\partial^2}{\partial t^2}v(t) = -Kv(t) \qquad\qquad \Delta_x u(x) = -Ku(x).$$

Die zeitunabhängige Differentialgleichung entspricht der Helmholtz-Gleichung.

Die Helmholtz-Gleichung kann auch auf andere Weise gewonnen werden. Die stationäre, also zeitunabhängige, Schrödinger-Gleichung[iii]

$$\frac{\hbar^2}{2m}\Delta\phi(x) - V(x)\phi(x) + E\phi(x) = 0$$

mit dem reduzierten Planck'schen Wirkungsquantum[iv] $\hbar^2$, der Masse m, der potentiellen Energie $V(x)$ und der Gesamtenergie E kann ebenfalls als Helmholtz-Gleichung aufgefasst werden

$$\Delta\phi(x) + \frac{2m}{\hbar^2}(E - V(x))\phi(x) = 0,$$

[iii] Trägt den Namen des österreichischen Physikers und Nobelpreisträgers Erwin Rudolf Josef Alexander Schrödinger, der die Gleichung 1926 formulierte.

[iv] Benannt nach dem deutschen Physiker und Nobelpreisträger Max Karl Ernst Ludwig Planck, der zeitweise auch an der Kieler Universität lehrte.

was einen quantenmechanischen Zugang darstellt.
Eine Lösung der Helmholtz-Gleichung (1.1.1) kann mit Hilfe der *Fundamentallösung* des Operators $-\Delta - \kappa^2$ konstruiert werden. Im dreidimensionalen Raum ist die Fundamentallösung durch

$$u_\kappa(x,y) := \frac{e^{\mathrm{i}\kappa\|x-y\|_2}}{4\pi\|x-y\|_2} \qquad \text{für fast alle } x,y \in \mathbb{R}^3 \tag{1.1.3}$$

gegeben [13, S. 198, S. 203]. Physikalisch entspricht diese Funktion einer Kugelwelle mit Zentrum in $y \in \mathbb{R}^3$. Die Fundamentallösung besitzt offensichtlich eine Singularität in $x = y$, somit ist sie nicht auf dem gesamten Gebiet differenzierbar. Auch ebene Wellen allein, die bei der hier vorgestellten Methode eine tragende Rolle einnehmen werden, können die Helmholtz-Gleichung lösen.

Zu Beginn wurde bereits erwähnt, dass es um ein Verfahren für die Helmholtz-Gleichung im *hochfrequenten* Bereich gehen soll. Dazu gilt es, zu klären, was hochfrequent in diesem Zusammenhang bedeutet. Es gibt keine einheitliche Definition, ab wann ein Problem mit $\kappa > 0$ als hochfrequent gilt, jedoch erscheint es sinnvoll, dies abhängig von der Gebietsgröße zu entscheiden. Entsprechend wird ein Problem dann als hochfrequent bezeichnet, wenn die Periodendauer im Vergleich zur Gebietsgröße gering ist. In der Praxis wird nicht das gesamte Gebiet auf einmal betrachtet, sondern meist eine Triangulation als Diskretisierung des Gebiets verwendet. Sei h der maximale Durchmesser eines Triangulationselements, dann beschreibt

$$\kappa h \gtrsim 1 \qquad \Longleftrightarrow \qquad h \gtrsim \frac{\lambda}{2\pi}$$

die Hochfrequenzbedingung. Um den Aufwand beim Lösen nicht zu sehr in die Höhe zu treiben, hält sich in der Praxis der Ansatz, mit mindestens fünf (teilweise zehn) Elementen der Triangulation pro Wellenlänge λ zu arbeiten. Dieser Ansatz kann mit dem Zusammenhang der Wellenzahl und der Wellenlänge (1.1.2) zu

$$\lambda \sim 5h \qquad \Longleftrightarrow \qquad \kappa h \gtrsim \frac{2\pi}{5}$$

umgeformt werden. Demnach wird in der Praxis, wie auch in dieser Arbeit, im hochfrequenten Bereich mit

$$\kappa h \sim 1,3 \qquad (\text{ für 10 Elemente } \kappa h \sim 0,6)$$

gearbeitet.

1.2 Problemstellungen

Dieser Abschnitt geht kurz auf einige gängige Problemstellungen der Helmholtz-Gleichung ein und orientiert sich hauptsächlich an [22] und [17, S. 90-102].

Um überhaupt eine vollständige Problemstellung zu haben, werden ein offenes Gebiet $\Omega_i \subset \mathbb{R}^3$, sein Rand $\Gamma := \partial\Omega_i$ und das mit $\Omega_a := \mathbb{R}^3 \setminus (\Omega_i \cup \Gamma) \subset \mathbb{R}^3$ bezeichnete Äußere des Gebiets benötigt, schematisch in Abbildung 1.1 gezeigt.

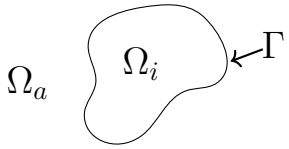
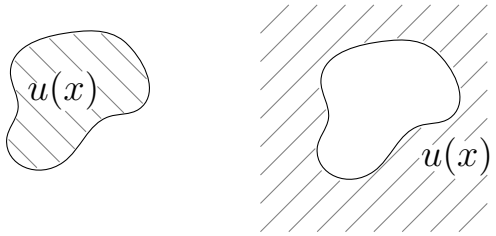


Abbildung 1.1: Geometrie

Grundsätzlich lassen sich drei Typen von Problemen, die jeweils mit unterschiedlichen Randwerten ausgestattet sein können, unterscheiden. Teilweise besteht auch die Möglichkeit, Randwerte direkt mit physikalischen Eigenschaften des Gebiets zu assoziieren.

Beim *Innenraumproblem* wird die Lösung der Differentialgleichung im Inneren und somit auf dem beschränkten Gebiet Ω_i gesucht, siehe Abbildung 1.2 (a). Ein mögliches Beispiel für diesen Problemtyp ist die Ausbreitung von akustischen Wellen im Inneren eines Fahrzeugs.
Entsprechend wird beim *Außenraumproblem* die Lösung der Differentialgleichung außerhalb des Gebiets, also in Ω_a, gesucht, siehe Abbildung 1.2 (b). Hier wäre die Ortung per Sonar ein denkbarer Anwendungsfall.



(a) Innenraumproblem (b) Außenraumproblem

Abbildung 1.2: Innen- und Außenraumproblem

Durch Kombination der beiden eben genannten Problemtypen ergibt sich der letzte Typ, beim *Ganzraumproblem* wird entsprechend die Lösung auf ganz $\mathbb{R}^3$ gesucht.
Zur Beschreibung des Verhaltens der Lösung auf dem Rand wird teilweise die *Normalenableitung*, also die Richtungsableitung entlang eines Normalenvektors, verwendet [33, S. 16]. Für die Definition der Normalenableitung ist wiederum ein zum Bildraum der betrachteten Funktion passendes Skalarprodukt notwendig.

Definition 1.1 (Komplexes Skalarprodukt)
Für $d \in \mathbb{N}$ definiere das komplexe Skalarprodukt *durch*

$$\langle x, y\rangle_{\mathbb{C}} := \sum_{i=1}^{d} x_i \overline{y_i} \qquad \text{für alle } x, y \in \mathbb{C}^d,$$

falls $x, y \in \mathbb{R}^d$ gilt, entspricht dies dem euklidischen Skalarprodukt[v].

Definition 1.2 (Normalenableitung)
Seien $d \in \mathbb{N}$, ein hinreichend glatt berandetes Gebiet $D \subset \mathbb{R}^d$ sowie eine auf D stetig differenzierbare Funktion f gegeben. Bezeichne mit $n(x)$ den für fast alle $x \in \partial D$ existierenden nach außen weisenden Normaleneinheitsvektor im Punkt x. Dann ist die Normalenableitung *von f auf dem Rand ∂D definiert durch*

$$\tfrac{\partial}{\partial n(x)} f(x) = \langle \nabla f(x), n(x)\rangle_{\mathbb{C}} \qquad \text{für fast alle } x \in \partial D,$$

wobei ∇f der Gradient von f sei.

Die später betrachteten Lipschitz-Gebiete[vi] haben hinreichend glatte Ränder, so dass für fast alle Punkte auf dem Rand der Normaleneinheitsvektor existiert.

1.2.1 Innenraumproblem

Das beim Innenraumproblem betrachtete Gebiet Ω_i ist sowohl offen als auch beschränkt. Um eine eindeutige Lösung zu erhalten, sind zusätzlich Randwerte nötig.
Die simpelsten Randwerte ergeben sich aus einer Forderung an das Verhalten der Lösung auf dem Rand Γ selbst, hierbei handelt es sich um *Dirichlet-Randwerte*[vii]. Das Dirichlet-Innenraumproblem ist damit durch

$$\begin{aligned} \Delta u(x) + \kappa^2 u(x) &= 0 && \text{für } x \in \Omega_i, \\ u(x) &= g_D(x) && \text{für } x \in \Gamma \end{aligned}$$

beschrieben. Im einfachsten Fall der homogenen Dirichlet-Randwerte ist dies durch die Funktion $g_D \equiv 0$ gegeben, dieser Spezialfall entspricht *schallweichen* Bedingungen in der Akustik. Ein Objekt wird als schallweich bezeichnet, falls eine akustische Welle ungehindert das Objekt und damit den Rand Γ passieren kann [14, S. 58].
Wird hingegen eine Forderung an das Verhalten der Normalenableitung der Lösung auf dem

[v] Namensgebend ist der antike griechische Mathematiker Euklid von Alexandria.
[vi] Benannt nach dem deutschen Mathematiker und Entwickler des Begriffs Rudolf Otto Sigismund Lipschitz.
[vii] Benannt nach dem deutschen Mathematiker Johann Peter Gustav Lejeune Dirichlet.

Rand gestellt, handelt es sich um *Neumann-Randwerte*[viii] und das Problem nimmt die Form

$$\begin{aligned} \Delta u(x) + \kappa^2 u(x) &= 0 && \text{für } x \in \Omega_i, \\ \tfrac{\partial}{\partial n(x)} u(x) &= g_N(x) && \text{für } x \in \Gamma \end{aligned}$$

an. Aus der Perspektive der Akustik führen homogene Neumann-Randwert zu *schallharten* Bedingungen. Eine, das Objekt treffende, Welle wird demnach vollständig zurückgeworfen, sie kann den Rand nicht passieren und somit auch nicht in den Außenraum dringen [14, S. 58].
Eine Variante, um komplexere Randwerte zu definieren, sind *Robin-Randwerte*[ix]. Sie erlauben eine Modellierung komplizierterer Objekte, indem sie Bedingungen an die Lösung selbst und ihre Normalenableitung *zugleich* stellen, was sie auch von gemischten Randwerten, welche für einzelne Teilgebiete von Γ entweder Dirichlet- oder Neumann-Randwerte vorgeben, unterscheidet. Ein Innenraumproblem dieser Bauart besitzt die Form

$$\begin{aligned} \Delta u(x) + \kappa^2 u(x) &= 0 && \text{für } x \in \Omega_i, \\ \tfrac{\partial}{\partial n(x)} u(x) + \mathrm{i}\alpha(x) u(x) &= g_R(x) && \text{für } x \in \Gamma, \end{aligned}$$

mit einer ortsabhängigen Funktion α. Physikalisch beschreibt α die Permeabilität der Oberfläche bezüglich der betrachteten Wellen. Problemstellungen dieser Art sind auch als *Impedanz-Problem* bekannt [17, S. 101].

1.2.2 Außenraumproblem

Für die Randwerte des Außenraumproblems gibt es dieselben Möglichkeiten wie beim Innenraumproblem. Außenraumprobleme sind jedoch auf einem unbeschränkten Gebiet Ω_a formuliert. Um in diesem Fall physikalisch sinnvolle und eindeutige Lösungen zu erhalten, muss eine zusätzliche Bedingung an die Lösung gestellt werden. Das Problem liegt darin begründet, dass mathematisch sowohl eine sich vom Objekt ins Unendliche ausbreitende als auch eine aus dem Unendlichen kommende und zum Objekt hinlaufende Welle Lösungen der Helmholtz-Gleichung im Außenraum darstellen. Die zweitgenannte Variante ist jedoch physikalisch unsinnig und es gilt, sie auszuschließen. Eine reale Welle klingt gedämpft durch physikalische Prozesse im Raum nach und nach ab. Diese Eigenschaft soll auch an die Lösungen der Helmholtz-Gleichung weiter gegeben werden, weshalb eine *Ausstrahlungs-* oder *Abklingbedingung* benötigt wird. Die erste für die Helmholtz-Gleichung verwendete ist die *Sommerfeld'sche Ausstrahlungsbedingung*[x]. Die hier definierte Ausstrahlungsbedingung bezieht sich auf die Helmholtz-Gleichung im Dreidimensionalen und ist so formuliert, dass sie einfallende Wellen der Bauart $\frac{e^{-\mathrm{i}\kappa\|x\|_2}}{4\pi\|x\|_2}$ herausfiltert [13, S. 216 ff.].

[viii] Benannt nach dem ungarisch-US-amerikanischen Mathematiker Jànos von Neumann.
[ix] Benannt nach dem französischen Mathematiker Victor Gustave Robin.
[x] Der deutsche Mathematiker und Physiker Arnold Sommerfeld führte in unendlichen Gebieten eine Ausstrahlungsbedingung ein, welche die physikalisch sinnvolle Lösung aus den mathematisch korrekten herausfil-

Definition 1.3 (Sommerfeld'sche Ausstrahlungsbedingung)
Die Sommerfeld'sche Ausstrahlungsbedingung an eine Lösung u der Helmholtz-Gleichung mit Wellenzahl κ und $r := \|x\|_2$ ist durch

$$\lim_{r\to\infty} r\left(\frac{\partial \tilde{u}}{\partial r} - \mathrm{i}\kappa\tilde{u}\right) = 0$$

gegeben, wobei $\tilde{u}$ die in Kugelkoordinaten definierte Funktion u sei, entsprechend ist $\frac{\partial}{\partial r}$ die partielle Ableitung entlang der radialen Koordinate.

Folglich sind die Außenraumprobleme mit jeweils einer der drei vorgestellten Bedingung für die Randwerte charakterisiert durch

$$\begin{aligned}
\Delta u(x) + \kappa^2 u(x) &= 0 && \text{für } x \in \Omega_a, && \\
u(x) &= g_D(x) && \text{für } x \in \Gamma, && \text{(Dirichlet-Problem)} \\
\tfrac{\partial}{\partial n(x)} u(x) &= g_N(x) && \text{für } x \in \Gamma, && \text{(Neumann-Problem)} \\
\tfrac{\partial}{\partial n(x)} u(x) + \mathrm{i}\alpha(x)u(x) &= g_R(x) && \text{für } x \in \Gamma, && \text{(Impedanz-Problem)} \\
\lim_{r\to\infty} r\left(\tfrac{\partial \tilde{u}}{\partial r} - \mathrm{i}\kappa\tilde{u}\right) &\to 0 && \text{für } r := \|x\|_2. &&
\end{aligned}$$

Das Erfüllen einer zusätzlichen Ausstrahlungsbedingung zu fordern, ist nicht die einzige Möglichkeit, eine eindeutige Lösung zu gewinnen. Da beim approximativen Lösen auf dem Computer keine unendlichen Gebiete dargestellt werden können, ist eine verbreitete Variante das Hinzufügen von künstlichen semipermeablen Rändern, die Wellen nur in einer Richtung passieren lassen.
Eine besondere Form der Außenraumprobleme sind die sogenannten *Streuprobleme*. Hierbei wird eine Welle auf das Objekt geschickt und es gilt, die vom Rand des Gebiets zurückgeworfene Welle zu bestimmen. Dazu wird die Lösung in zwei Teile gesplittet

$$u(x) = u_i(x) + u_s(x) \qquad \text{für } x \in \Omega_a,$$

eine bekannte einfallende Welle u_i und den unbekannten reflektierten Teil u_s.

1.2.3 Existenz und Eindeutigkeit der Lösung

Ein weiterer wichtiger Aspekt ist die Frage nach der Existenz und Eindeutigkeit der Lösung der Helmholtz-Gleichung.

tert. Die ursprüngliche Formulierung besteht aus der *Ausstrahlungs-* und einer zusätzlichen *Endlichkeitsbedingung* $\lim_{\|x\|_2\to\infty} u(x) \neq \infty$ [47, S. 327 ff.]. Rellich zeigte 1940, dass für reelle κ die Ausstrahlungs- die Endlichkeitsbedingung impliziert und führte eine Integralformulierung der Ausstrahlungsbedingung $\lim_{r\to\infty} \int_{\|x\|_2=r} |u(x)|^2 dx = 0$, welche für Randintegrale ausreichend ist, ein [42].

Unabhängig von den konkret gewählten Randwerten hängt die Eindeutigkeit der Lösung der Helmholtz-Gleichung vom Rand Γ des Gebiets selbst und der Wellenzahl κ ab. Folglich gestaltet es sich schwieriger, allgemeine Aussagen zur Existenz und Eindeutigkeit der Lösung zu treffen.
Angaben zur Existenz und Eindeutigkeit für das Innenraumproblem finden sich unter anderem in [38, S. 286],[13, S. 235-237], welche hier ohne Beweis zitiert werden sollen.

Satz 1.4 (Eindeutigkeit Innenraumproblem)
Es existiert eine aufsteigende Folge von positiven Wellenzahlen $\{\kappa_i\}_{i\in\mathbb{N}}$, für die das Innenraumproblem keine eindeutige Lösung besitzt, für alle anderen Wellenzahlen $\kappa > 0$ existiert sowohl für das Dirichlet- als auch das Neumann-Problem genau eine Lösung. Für den Fall $\kappa = 0$, also im Fall der Laplace-Gleichung, ist das Dirichlet-Problem eindeutig lösbar, während das Neumann-Problem nur bis auf eine additive Konstante eindeutig bestimmt ist.

Die Folge der positiven Wellenzahlen besteht aus den Wurzeln der Eigenwerte für das zur Helmholtz-Gleichung gehörende Laplace-Problem mit Nullrandbedingungen für Dirichlet- oder Neumann-Randwerte. Das Superpositionsprinzip liefert dann durch das Addieren der entsprechenden Eigenfunktionen unendlich viele Lösungen der Helmholtz-Gleichung.
Für Robin-Randwerte macht [17, S. 103 ff.] eine Aussage zur Existenz und Eindeutigkeit unter zusätzlichen Forderungen an α.
Stärkere Eindeutigkeitsaussagen sind für komplexe Wellenzahlen möglich.

Bemerkung 1 (Komplexe Wellenzahlen): *Für alle Wellenzahlen $\kappa \in \mathbb{C}$ mit $\Im(\kappa) > 0$ existiert sowohl für das Dirichlet- als auch für das Neumann-Innenraumproblem höchstens eine Lösung [13, S. 237].*

Aussagen für das Außenraumproblem können in [17, S. 106][13, S. 232] gefunden werden.

Satz 1.5 (Eindeutigkeit Außenraumproblem)
Sowohl das Dirichlet- als auch das Neumann-Außenraumproblem besitzen höchstens eine Lösung, für das Dirichlet-Außenraumproblem existiert exakt eine Lösung.

Unter der Bedingung, dass $\Re(\alpha) \geq 0$ können die Existenz- und Eindeutigkeitsaussagen des Dirichlet-Außenraumproblems auch für das Impedanz-Problem gezeigt werden [17, S. 106].
Wie im Fall des Innenraumproblems kann die Aussage auch für komplexe Wellenzahlen modifiziert werden.

Bemerkung 2 (Komplexe Wellenzahlen): *Für alle Wellenzahlen $\kappa \in \mathbb{C}$ mit $\Im(\kappa) > 0$ existiert sowohl für das Dirichlet- als auch für das Neumann-Außenraumproblem höchstens eine Lösung [13, S. 232].*

1.3 Randelementmethode

Der Grundgedanke der *Randelementmethode* BEM (engl. *boundary element method*) ist es, den Prozess des Lösens vom gesamten d-dimensionalen Gebiet Ω auf den $(d-1)$-dimensionalen Rand Γ zu reduzieren. Im Fall der dreidimensionalen Helmholtz-Gleichung bedeutet dies die Reduktion von einem Volumen auf eine Oberfläche.
Zwar ermöglicht die Reduktion der Problemgröße, dass unendliche Gebiete bearbeitet werden können, sie hat jedoch auch ihren Preis. Es muss in Kauf genommen werden, dass eine durch eine Finite-Elemente-Diskretisierung des Gebiets sehr große, aber dafür schwach besetzte Matrix bei den Randelementmethoden deutlich kleiner, aber dafür vollbesetzt sein wird. Zusätzlich erfordert die Randelementmethode eine Umformulierung, so dass eine Betrachtung des Rands, auf dem die Differentialgleichung zunächst gar nicht definiert ist, ausreicht.
In diesem Kapitel werden zu Beginn die nötigen Grundlagen aus der Funktionalanalysis angerissen, die Anwendung findenden Potentiale vorgestellt und anschließend an einem kurzen Beispiel vorgeführt, wie mit Hilfe der Potentiale das Problem in ein lineares Gleichungssystem überführt wird.

1.3.1 Funktionalanalytische Grundlagen

Um die Randelementmethode anwenden zu können, ist es unter anderem notwendig, schwache Ableitungen zu betrachten, die dazugehörigen Sobolev-Räume[xi] einzuführen sowie die zulässigen Gebiete einzuschränken.
Für eine ausführliche Einführung und allgemeine theoretische Aussagen sei der Leser zum Beispiel auf [21] oder [38] verwiesen.

Für beliebige Gebiete und Ränder sind keine Aussagen möglich, entsprechend sind gewisse Regularitätsbedingungen für den Rand Γ notwendig. Aus diesem Grund wird die Theorie auf *Lipschitz-Gebiete* eingeschränkt [38, S. 89 ff.], [49, S. 21].

Definition 1.6 (Lipschitz-Gebiet)
Bezeichne eine offene Menge $\Omega \subset \mathbb{R}^3$ als Lipschitz-Gebiet*, falls ihr Rand $\Gamma := \partial\Omega$ kompakt ist, ein $J \in \mathbb{N}$ sowie endliche Familien von Teilmengen $\{U_j\}_{j\in\mathbb{N}_{<J}}$, $\{\Omega_j\}_{j\in\mathbb{N}_{<J}}$ des $\mathbb{R}^3$ und eine endliche Menge von differenzierbaren Funktionen $\{\xi_j\}_{j\in\mathbb{N}_{<J}}$, die von Teilmengen $\Omega'_j \subset \mathbb{R}^2$ in den $\mathbb{R}$ abbilden, existieren mit*

[xi] Sie sind nach dem russischen Mathematiker Sergei Lwowitsch Sobolew benannt, der im 20. Jahrhundert die Theorie der schwach differenzierbaren Funktionen und ihrer Räume weiter entwickelte.

i. *für alle $j \in \mathbb{N}_{<J}$ existiert ein $L_j \in \mathbb{R}_{\geq 0}$, so dass ξ_j Lipschitz-stetig ist*

$$|\xi_j(x) - \xi_j(y)| \leq L_j \|x - y\|_2 \qquad \text{für alle } x, y \in \Omega'_j,$$

ii. *für alle $j \in \mathbb{N}_{<J}$ ist U_j offen und es gilt $\Gamma \subset \bigcup_{j \in \mathbb{N}_{<J}} U_j$, entsprechend bildet die Familie $\{U_j\}_{j \in \mathbb{N}_{<J}}$ eine endliche offene Überdeckung von Γ,*

iii. *für alle $j \in \mathbb{N}_{<J}$ gilt $U_j \cap \Omega = U_j \cap \Omega_j$ und*

iv. *für alle $j \in \mathbb{N}_{<J}$ kann Ω_j in einem lokalen Koordinatensystem auf einen Lipschitz-Hypographen überführt werden. Das heißt, es existiert eine Komposition ι_j aus einer Rotation und einer Translation mit*

$$\iota_j(\Omega_j) = \left\{ x \in \mathbb{R}^3 \mid x' \in \Omega'_j,\, x_3 < \xi_j(x') \right\},$$

wobei x' der Vektor sei, der nur aus den ersten beiden Komponenten von x besteht, also $x' = (x_1, x_2)$.

Definiere zur Überdeckung passende Teilstücke des Rands

$$\Gamma_j := U_j \cap \Gamma \qquad \text{für alle } j \in \mathbb{N}_{<J},$$

so dass der Rand auch mit

$$\Gamma = \bigcup_{j \in \mathbb{N}_{<J}} \Gamma_j$$

geschrieben werden kann. Die einzelnen Teilstücke Γ_j können auch als Graph der Lipschitz-stetigen Funktionen ξ_j in dem lokalen Koordinatensystem dargestellt werden, so dass

$$\iota_j(\Gamma_j) = \left\{ x \in \iota_j(U_j) \mid x' \in \Omega'_j, x = (x', \xi_j(x')) \right\} \qquad \text{für alle } j \in \mathbb{N}_{<J}$$

gilt.
Da ι_j als Komposition einer Rotation und einer Translation invertierbar ist, kann Γ_j zu jedem $j \in \mathbb{N}_{<J}$ mit der Teilmenge $\Omega'_j \in \mathbb{R}^2$ durch

$$\Gamma_j = \left\{ x \in U_j \mid y \in \Omega'_j,\, x = \iota_j^{-1}(y, \xi_j(y)) \right\}$$

dargestellt werden. Entsprechend lässt sich eine lokale Parametrisierung des Rands finden, so dass zu jedem $j \in \mathbb{N}_{<J}$ eine Funktion

$$\chi_j : \Omega'_j \to \mathbb{R}^3, \qquad \chi_j(y) = \iota_j^{-1}(y, \xi_j(y)) \tag{1.3.1}$$

existiert.

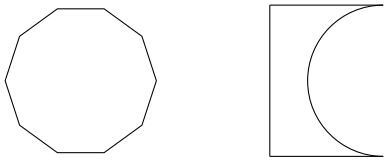


Abbildung 1.3: Beispiel und Gegenbeispiel für Lipschitz-Gebiete

Für die meisten Anwendungen sind Lipschitz-Gebiete hinreichend allgemein. Alle Polyeder sind Lipschitz-Gebiete und Triangulationen von realen Gegenständen werden aus Polyedern gebildet. Die Abbildung 1.3 zeigt links ein Beispiel für ein Lipschitz-Gebiet und rechts ein Gegenbeispiel, das aufgrund der Spitzen die Eigenschaften nicht erfüllt.
Innenräume Ω_i werden durch beschränkte Lipschitz-Gebiete dargestellt, während der dazugehörige Außenraum Ω_a ein unbeschränktes Lipschitz-Gebiet ist, beide teilen sich den Lipschitz-Rand Γ [38, S. 90 f.].

Definition 1.7 (Raum der differenzierbaren Funktionen)
Für ein $0 \leq p < \infty$ und eine offene reelle Menge Ω bezeichne mit $C^p(\Omega)$ die Menge aller komplexwertigen p-mal stetig differenzierbaren Funktionen auf Ω. Entsprechend bezeichne durch $C^\infty(\Omega)$ den Raum der unendlich oft differenzierbaren Funktionen auf Ω sowie $C_0^\infty(\Omega)$ die Teilmenge der Funktionen aus $C^\infty(\Omega)$ mit kompaktem Träger.

Um die starke Forderung der Differenzierbarkeit abzuschwächen, sind zunächst Lebesgue-messbare[xii] Räume nötig, mit denen die Räume der *schwach differenzierbaren* Funktionen eingeführt werden können. Die Lebesgue-Räume lassen sich mit einer passenden Norm ausstatten und im Spezialfall auch zu Hilbert-Räumen[xiii] erweitern [38, S. 58 ff.][21, S. 71].

Definition und Lemma 1.8 (Lebesgue-Räume)
Für eine messbare Menge $\Omega \subset \mathbb{R}^3$ und jedes $1 \leq p < \infty$ bezeichne mit $L^p(\Omega)$ den Raum aller Äquivalenzklassen, welche durch die Äquivalenzrelation

$$f \sim g \iff f = g \qquad \text{fast überall auf } \Omega$$

definiert sind, Lebesgue-messbarer Funktionen f mit

$$\int_\Omega |f(x)|^p \, \mathrm{d}x < \infty.$$

*Für alle $1 \leq p < \infty$ definiere die dazugehörige L^p-*Norm *durch*

$$\|f\|_{L^p(\Omega)} := \left(\int_\Omega |f(x)|^p \, \mathrm{d}x \right)^{1/p} \qquad \text{für alle } f \in L^p(\Omega).$$

[xii] Benannt nach dem französischen Mathematiker und Begründer der Maßtheorie Henri Léon Lebesgue.
[xiii] Benannt nach dem deutschen Mathematiker David Hilbert.

Mit dem Skalarprodukt

$$\langle f, g\rangle_{L^2(\Omega)} := \int_\Omega \langle f(x), g(x)\rangle_{\mathbb{C}}\, \mathrm{d}x \qquad \text{für alle } f, g \in L^2(\Omega)$$

bildet $L^2(\Omega)$ einen Hilbert-Raum.

Um die Notation partieller Ableitungen zu vereinfachen, nutze *Multiindizes* und führe mit ihrer Hilfe schwache Ableitungen ein [21, S. 88 ff.].

Definition 1.9 (Multiindex und dazugehörige Ableitung)
Für $d \in \mathbb{N}$ bezeichne $\alpha \in \mathbb{N}_0^d$ mit $\alpha = (\alpha_1, \ldots, \alpha_d)$ als Multiindex. *Der Betrag eines Multiindex ist gegeben durch*

$$|\alpha| := \sum_{i=1}^{d} \alpha_i \qquad \text{für alle } \alpha \in \mathbb{N}_0^d.$$

Definiere die α-te Ableitung einer hinreichend oft differenzierbaren Funktion f mit

$$\partial^\alpha f := \partial_1^{\alpha_1} \partial_2^{\alpha_2} \cdots \partial_d^{\alpha_d} f.$$

Definition 1.10 (Schwache Ableitung)
Sei ein Multiindex $\alpha \in \mathbb{N}_0^d$ gegeben. Eine Funktion $f \in L^2(\Omega)$ besitzt eine α-te schwache Ableitung auf Ω, falls ein $f_\alpha \in L^2(\Omega)$ existiert mit

$$\int_\Omega f(x)\partial^\alpha \phi(x)\, \mathrm{d}x = (-1)^{|\alpha|} \int_\Omega f_\alpha(x)\phi(x)\, \mathrm{d}x \qquad \text{für alle } \phi \in C_0^\infty(\Omega).$$

Bezeichne f_α dann als α-te schwache Ableitung von f.

Die Definition ist durch die partielle Integration motiviert. Natürlich kann eine schwache Ableitung auch auf L^p-Räumen für $p \in \mathbb{N}$ definiert werden, die klassische Formulierung wäre für $L^1_{loc}(\Omega)$. Bei $L^p_{loc}(\Omega)$ handelt es sich um den Raum der *lokal integrierbaren Funktionen*, also aller Funktionen, die auf jeder offenen Teilmenge $U \subset \Omega$, mit $\overline{U}$ ist kompakt, in $L^p(U)$ enthalten sind.

Lemma 1.11 (Eindeutigkeit)
Sofern die schwache Ableitung existiert, ist sie eindeutig. Falls eine klassische Ableitung existiert, so stimmt sie mit der schwachen überein. [21, Lem. 5.4].

Eindeutigkeit bedeutet aufgrund der Äquivalenzklassenbildung, dass, falls zwei schwache Ableitungen existieren, diese bis auf einer Nullmenge übereinstimmen.

Definition 1.12 (Schwach differenzierbar)
Schreibe die α-te schwache Ableitung f_α einer Funktion $f \in L^2(\Omega)$ mit $\partial^\alpha f := f_\alpha$. Falls zu einer Funktion $f \in L^2(\Omega)$ und einem $m \in \mathbb{N}$ alle α-ten schwachen Ableitungen mit $|\alpha| \leq m$ existieren, bezeichne die Funktion f als m-mal schwach differenzierbar.

Mit Hilfe des Konzepts der schwachen Differenzierbarkeit kann ein entsprechender Funktionsraum definiert werden [21, S. 91].

Definition und Lemma 1.13 (Sobolev-Raum)
Für ein Lipschitz-Gebiet Ω und $m \in \mathbb{N}_0$ bezeichne den Raum aller m-mal schwach differenzierbaren Funktionen $f \in L^2(\Omega)$ als Sobolev-Raum $H^m(\Omega)$. *Versehe den Raum $H^m(\Omega)$ mit der* Sobolev-Norm

$$\|f\|_{H^m(\Omega)} = \left(\sum_{|\alpha| \leq m} \|\partial^\alpha f\|^2_{L^2(\Omega)} \right)^{1/2} \qquad \text{für alle } f \in H^m(\Omega).$$

Der Sobolev-Raum $H^m(\Omega)$ bildet einen Hilbert-Raum mit dem Skalarprodukt

$$\langle f, g \rangle_{H^m(\Omega)} = \sum_{|\alpha| \leq m} \langle \partial^\alpha f, \partial^\alpha g \rangle_{L^2(\Omega)} \qquad \text{für alle } f, g \in H^m(\Omega).$$

Die Potentiale erfordern eine Verallgemeinerung der Sobolev-Räume $H^s(\Omega)$ auf reelle Indizes zunächst mit $s \geq 0$ [21, S. 124].

Definition und Lemma 1.14 (Gebrochener Sobolev-Raum)
Seien ein Lipschitz-Gebiet $\Omega \subset \mathbb{R}^3$ sowie ein reelles nicht ganzzahliges $s > 0$ gegeben. Schreibe $s = m + \sigma$ mit $m \in \mathbb{N}_0$ und $\sigma \in (0, 1)$ und definiere die Norm für alle $f \in L^2(\Omega)$ mit

$$\|f\|_{H^s(\Omega)} := \left(\|f\|^2_{H^m(\Omega)} + \sum_{|\alpha|=m} |\partial^\alpha f|^2_{\sigma,\Omega} \right)^{1/2},$$

wobei $|\cdot|_{\sigma,\Omega}$ eine Halbnorm mit

$$|f|^2_{\sigma,\Omega} := \int_\Omega \int_\Omega \frac{|f(x)-f(y)|^2}{|x-y|^{3+2\sigma}} \, \mathrm{d}x \, \mathrm{d}y$$

ist. Der Raum $H^s(\Omega)$ ist durch die Funktionen $f \in H^m(\Omega)$ mit endlicher Norm $\|f\|_{H^s(\Omega)}$ gegeben. Der Raum $H^s(\Omega)$ bildet mit dem Skalarprodukt

$$\langle f, g \rangle_{H^s(\Omega)} := \langle f, g \rangle_{H^m(\Omega)} + \sum_{|\alpha|=m} \langle \partial^\alpha f, \partial^\alpha g \rangle_{H^\sigma(\Omega)} \qquad \text{für alle } f, g \in H^s(\Omega),$$

wobei $\langle \cdot, \cdot \rangle_{H^\sigma(\Omega)}$ durch

$$\langle f, g \rangle_{H^\sigma(\Omega)} := \int_\Omega \int_\Omega \frac{(f(x)-f(y))\overline{(g(x)-g(y))}}{|x-y|^{3+2\sigma}} \, \mathrm{d}x \, \mathrm{d}y \qquad \text{für alle } f, g \in H^s(\Omega)$$

gegeben ist, einen Hilbert-Raum.

Die Erweiterung auf negative reelle Indizes s erfolgt über die Dualräume [21, S. 122 ff.], [38, S. 75 ff.].
Bezeichne mit $H^{-s}(\Omega)$ für $s \in \mathbb{R}_{>0}$ den Dualraum zu $H^s(\Omega)$, also den Raum aller stetigen linearen Funktionale von $H^s(\Omega)$. Definiere die Operatornorm auf $H^{-s}(\Omega)$ mit Hilfe der *dualen Paarung* $\langle \cdot, \cdot \rangle$ durch

$$\|f\|_{H^{-s}(\Omega)} = \sup \left\{ \tfrac{|\langle f,g \rangle|}{\|g\|_{H^s(\Omega)}} \mid g \in H^s(\Omega), g \neq 0 \right\}.$$

Die Räume $H^s(\Omega)$ und $H^{-s}(\Omega)$ sind isometrisch isomorph[xiv] zu einander.
Ist die Menge Ω beschränkt, gilt für alle $-\infty < s < t < \infty$ die ursprünglich von Rellich gezeigte Einbettung $H^t(\Omega) \subset H^s(\Omega)$ [38, Thm. 3.27].

Zu guter Letzt ist es nötig, Sobolev-Räume auch für den Rand des Gebiets Γ zu definieren [38, S. 96 ff.], [49, S. 21]. Dies kann durch Zurückführen des Maßes auf die Lebesgue-Maße für die einzelnen Randstücke Γ_j im lokalen Koordinatensystem geschehen. Verwende dazu eine zur endlichen Überdeckung des Rands Γ passende Partition der Eins aus den Funktionen $\{\phi_j\}_{j \in \mathbb{N}_{<J}}$ aus $C^\infty(\mathbb{R}^3, \mathbb{R})$ mit kompaktem Träger, welche

- $\operatorname{supp}(\phi_j) \subset U_j$, $\quad 0 \leq \phi_j(x) \leq 1$ für alle $j \in \mathbb{N}_{<J}, x \in \mathbb{R}^3$
- $\sum\limits_{j \in \mathbb{N}_{<J}} \phi_j(x) = 1$ $\quad$ für alle $x \in \Gamma$

erfüllen, um jede Funktion $v(x) : \Gamma \to \mathbb{R}$ mit

$$v(x) = \sum_{j \in \mathbb{N}_{<J}} \phi_j(x) v(x) \qquad \text{für alle } x \in \Gamma$$

schreiben und damit globale Funktionen auf die Teilstücke der Überdeckung zurückführen zu können. Das Lebesgue-Maß von dem Teilstück Γ_j ist dann durch

$$|\Gamma_j| = \int_{\Omega'_j} \sqrt{1 + |\nabla \xi_j(y)|^2}\, \mathrm{d}y \tag{1.3.2}$$

gegeben [38, S. 97]. Entsprechend ergibt sich das Lebesgue-Integral mit der Parametrisierung (1.3.1)

$$\int_{\Gamma_j} f(x) dx = \int_{\Omega'_j} f(\chi_j(y)) \sqrt{1 + |\nabla \xi_j(y)|^2}\, \mathrm{d}y,$$

welches mit der Partition der Eins additiv auf den ganzen Rand Γ fortgesetzt werden

$$\int_\Gamma f(x) dx = \sum_{j \in \mathbb{N}_{<J}} \int_{\Gamma_j} \phi_j(x) f(x) dx$$

kann. Auf eine gesonderte Kennzeichnung des Oberflächenmaßes soll verzichtet werden, da aus dem Kontext heraus ersichtlich ist, wann es Anwendung findet.

[xiv] Zwei Räume sind genau dann isometrisch isomorph zueinander, wenn zwischen ihnen ein Vektorraumisomorphismus existiert, der ebenfalls ein Isometrie ist.

Definition 1.15 (Lebesgue-Raum auf dem Rand)
Für ein $j \in \mathbb{N}_{<J}$ ist der Lebesgue-Raum auf dem Rand $L^2(\Gamma_j)$ *mit dem Oberflächenmaß* (1.3.2) *analog zur Definition 1.8 definiert. Mit Hilfe der Partition der Eins können diese Lebesgue-Räume zu $L^2(\Gamma)$ fortgesetzt werden.*

Die Sobolev-Räume auf dem Rand werden auf Sobolev-Räume auf Teilmengen des $\mathbb{R}^2$ zurückgeführt [38, S. 98 ff.], [49, S. 21].

Definition und Lemma 1.16 (Sobolev-Raum auf dem Rand)
Seien ein Lipschitz-Gebiet $\Omega \subset \mathbb{R}^3$ und die lokalen Parametrisierungen $\{\chi_j\}_{j\in\mathbb{N}_{<J}}$ (1.3.1) *seines Rands gegeben. Sei $f_{\chi_j}(y) := f(\chi_j(y))$ für alle $j \in \mathbb{N}_{<J}$, $y \in \Omega'_j$ und $f \in L^2(\Gamma_j)$. Für $s \in [0,1]$ und $j \in \mathbb{N}_{<J}$ ist der* Sobolev-Raum auf dem Randstück Γ_j *dann definiert durch*

$$H^s(\Gamma_j) := \left\{ f \in L^2(\Gamma_j) \mid f_{\chi_j} \in H^s(\Omega'_j) \right\}.$$

Zusammen mit dem Skalarprodukt und der induzierten Norm für alle $f, g \in H^s(\Gamma_j)$

$$\langle f, g\rangle_{H^s(\Gamma_j)} := \langle f_{\chi_j}, g_{\chi_j}\rangle_{H^s(\Omega'_j)} \qquad \|f\|_{H^s(\Gamma_j)} := \left(\langle f, f\rangle_{H^s(\Gamma_j)}\right)^{\frac{1}{2}}$$

bildet $H^s(\Gamma_j)$ für alle $j \in \mathbb{N}_{<J}$ einen Hilbert-Raum.
Dies kann über die Partition der Eins und

$$\langle f, g\rangle_{H^s(\Gamma)} := \sum_{j\in\mathbb{N}_{<J}} \langle \phi_j f_{\chi_j}, \phi_j g_{\chi_j}\rangle_{H^s(\Gamma_j)} \qquad \text{für alle } f, g \in H^s(\Gamma)$$

zu einem Sobolev-Raum auf dem gesamten Rand $H^s(\Gamma)$*, der ebenfalls ein Hilbert-Raum ist, fortgesetzt werden. Für $\hat{s} \in [-1, 0)$ definiere $H^{\hat{s}}(\Gamma)$ als Dualraum von $H^{-\hat{s}}(\Gamma)$.*

Bemerkung 3 (Eindeutigkeit): *Unabhängig von der gewählten Überdeckung des Rands und der konkreten Partition der Eins ergibt sich derselbe Raum $H^s(\Gamma)$, jedoch nur mit einer äquivalenten Norm [38, S. 98 f.].*

1.3.2 Potentiale und Operatoren

Im Folgenden sei der Einfachheit halber das Gebiet Ω entweder Ω_i oder Ω_a, je nachdem, was für eine Problemstellung gerade betrachtet werden soll.
Beim Lösen der homogenen Helmholtz-Gleichung reicht es aus, Integrale auf dem Rand Γ zu bestimmen. Dies geschieht mit Hilfe von *Potentialen*. Die verschiedenen Typen von Potentialen haben gemeinsam, dass sie mit der Fundamentallösung (1.1.3) und mit deren Derivaten arbeiten. Direkt mit der Fundamentallösung agiert das *Einfachschichtpotential* [13, S. 207 ff.][48, S. 4].

Definition 1.17 (Einfachschichtpotential)
Das Einfachschichtpotential $V_\kappa : H^{-\frac{1}{2}}(\Gamma) \to H^1(\Omega)$ *ist für eine Dichte* w *durch*

$$V_\kappa[w](x) := \int_\Gamma u_\kappa(x,y)w(y)\,\mathrm{d}y \qquad \textit{für alle } x \in \mathbb{R}^3 \setminus \Gamma \tag{1.3.3}$$

gegeben, wobei $u_\kappa(x,y)$ *die Fundamentallösung* (1.1.3) *ist.*

Das Einfachschichtpotential an sich löst die Helmholtz-Gleichung [13, S. 207]. Es gilt

$$\Delta V_\kappa[w](x) + \kappa^2 V_\kappa[w](x) = \Delta_x \int_\Gamma u_\kappa(x,y)w(y)\,\mathrm{d}y + \kappa^2 \int_\Gamma u_\kappa(x,y)w(y)\,\mathrm{d}y.$$

Für $x \neq y$ ist die Fundamentallösung stetig differenzierbar nach x und x stammt je nach Formulierung aus Ω_i oder Ω_a. Die Integration und Differentiation dürfen vertauscht werden und die Integrale zusammengefasst. Dies führt zu

$$\Delta V_\kappa[w](x) + \kappa^2 V_\kappa[w](x) = \int_\Gamma w(y)\left(\Delta_x u_\kappa(x,y) + \kappa^2 u_\kappa(x,y)\right)\,\mathrm{d}y.$$

Da die Fundamentallösung die Helmholtz-Gleichung erfüllt, ist der Integrand 0 und es folgt

$$\Delta V_\kappa[w](x) + \kappa^2 V_\kappa[w](x) = 0.$$

Jedoch führt das Einfachschichtpotential nicht sofort auf die Lösung, sondern zunächst auf die sogenannte *Dichtefunktion* w, die so bestimmt werden muss, dass die gegebenen Randwerte erfüllt sind. Mit der Dichtefunktion kann im Anschluss die eigentliche Lösung $u(x)$ punktweise über die Beziehung $V_\kappa[w](x) = u(x)$ für $x \in \Omega$ bestimmt werden. Folglich wird das Lösen mit Hilfe des Einfachschichtpotentials auch als *indirekter Ansatz* bezeichnet.

Bemerkung 4 (Idee Einfachschichtpotential): *Wird eine Variable, zum Beispiel* y*, festgehalten (hier als festes* y_0*), so löst die Funktion*

$$g(x,y_0) = \frac{e^{i\kappa\|x-y_0\|_2}}{4\pi\|x-y_0\|_2}$$

die Helmholtz-Gleichung. Anstatt eines einzelnen y_0 *wird eine Menge von Punkten* $\{y_i\}_{i\in P}$ *betrachtet. Mit dem Superpositionsprinzip folgt, dass auch eine Linearkombination*

$$\tilde{u}_\kappa(x) = \sum_{i\in P} w_i \frac{e^{i\kappa\|x-y_i\|_2}}{4\pi\|x-y_i\|_2},$$

mit Gewichten w_i *die Helmholtz-Gleichung erfüllt. Das Einfachschichtpotential kann als Übergang von endlich zu unendlich vielen Punkten und damit von einer Summe zu einem Integral betrachtet werden [46, S. 8].*

Die Verwendung der Normalenableitung (siehe Definition 1.2) für die y-Koordinate der Fundamentallösung führt zum *Doppelschichtpotential* [13, S. 207 ff.][48, S. 4].

Definition 1.18 (Doppelschichtpotential)
Das Doppelschichtpotential $K_\kappa : H^{\frac{1}{2}}(\Gamma) \to H^1(\Omega)$ *ist für eine Dichte* v *gegeben durch*

$$K_\kappa[v](x) := \int_\Gamma v(y) \tfrac{\partial}{\partial n(y)} u_\kappa(x,y)\, \mathrm{d}y \qquad \textit{für alle } x \in \mathbb{R}^3 \setminus \Gamma,$$

wobei $u_\kappa(x,y)$ *die Fundamentallösung* (1.1.3) *ist.*

Ebenso wie das Einfachschichtpotential löst das Doppelschichtpotential die Helmholtz-Gleichung, der Beweis erfolgt auf die gleiche Weise wie beim Einfachschichtpotential, benötigt aber noch zusätzlich den Satz von Schwarz[xv], um die Reihenfolge der partiellen Ableitungen vertauschen zu dürfen [13, S. 207]. Auch das Anwenden des Doppelschichtpotentials stellt einen *indirekten Ansatz* dar, bei dem zunächst nur die Dichtefunktion v bestimmt wird.

Bemerkung 5 : *Für* $\kappa > 0$ *erfüllen sowohl das Einfach- als auch das Doppelschichtpotential die Sommerfeld'sche Ausstrahlungsbedingung [22, S. 35][13, S. 225].*

Schon in der Formulierung (1.3.3) bahnt sich eine Hürde an, denn es wird gefordert, dass $x \in \Omega$ ist, aber eben kein Element des Rands Γ, da dort die Fundamentallösung für $x = y$ singulär ist. Um die Anzahl der auftretenden Freiheitsgrade möglichst klein zu halten, soll die Berechnung aber gerade auf den Rand Γ eingeschränkt werden. Aus diesem Grund werden Grenzwerte der Potentiale auf dem Rand betrachtet, bei denen die zugehörigen Randwerte der Problemstellung berücksichtigt werden müssen. Hierfür werden *Spuroperatoren* verwendet und verschiedene *Spuren* unterschieden [22, S. 28].

Definition 1.19 (Dirichlet-Spur)
Sei eine Funktion $u \in C(\Omega)$ *gegeben. Wenn der Grenzwert in* $\tilde{x} \in \Gamma$ *existiert, ist die* Dirichlet-Spur γ *definiert durch*

$$\gamma u(\tilde{x}) = \lim_{\Omega \ni x \to \tilde{x}} u(x).$$

Werden Ω_i *und* Ω_a *betrachtet, bezeichne die Spur im Inneren als* innere Dirichlet-Spur *mit* γ^{int} *und die Spur im Äußeren als* äußere Dirichlet-Spur *mit* γ^{ext}*, soweit diese existieren.*

Ebenso lässt sich mit der Normalenableitung verfahren, die zur Neumann-Spur führt.
Für fast alle $x \in \Gamma$ mit $\hat{x} = \iota_j(x)$ und $\hat{x} = (\hat{x}', \hat{x}_3)$ existiert der äußere Normalenvektor $n(x)$ und kann mit der Hilfe der Lipschitz-stetigen Funktion ξ_j und der Rotation Q_j der

[xv] Benannt nach dem deutschen Mathematiker Hermann Amandus Schwarz.

Bewegung ι_j durch

$$n(x) = Q_j^*(\hat{n}(\hat{x})) \qquad \text{mit} \qquad \hat{n}(\hat{x}) = \frac{(-\nabla\xi_j(\hat{x}'),1)^T}{\sqrt{1+|\nabla\xi_j(\hat{x}')|^2}}$$

berechnet werden.

Definition 1.20 (Neumann-Spur)
Sei eine Funktion $u \in C^1(\Omega)$ gegeben. Wenn der Grenzwert für $\tilde{x} \in \Gamma$ existiert, ist die Neumann-Spur ∂_n *definiert durch*

$$\partial_n u(\tilde{x}) = \lim_{\Omega \ni x \to \tilde{x}} \langle \nabla u(x), n(\tilde{x}) \rangle_{\mathbb{C}},$$

wobei $n(\tilde{x})$ den äußeren Normalenvektor auf dem Rand Γ im Punkt $\tilde{x}$ bezeichne. Werden sowohl Ω_i als auch Ω_a betrachtet, bezeichne die Spur im Inneren als innere Neumann-Spur *mit ∂_n^{int} und die Spur im Äußeren als* äußere Neumann-Spur *mit ∂_n^{ext}, soweit diese existieren.*

Für den äußeren Normalenvektor bezüglich des Rands gilt die Konvention, dass dieser immer vom Inneren eines beschränkten Gebiets weg zeigt, was dazu führt, dass die Normalenvektoren für die innere und äußere Neumann-Spur in dieselbe Richtung zeigen.
Die oben nur für stetige beziehungsweise stetig differenzierbare Funktionen definierten Spuren können auf Sobolev-Räume fortgesetzt werden. Die *Spursätze* behandeln diese Fortsetzungen, um auf den Rändern der Gebiete arbeiten zu können.

Satz 1.21 (Spursatz Dirichlet-Spur)
Für alle $s \in \left(\frac{1}{2}, \frac{3}{2}\right)$ existiert eine eindeutige beschränkte Fortsetzung

$$\gamma : H^s(\Omega) \to H^{s-\frac{1}{2}}(\Gamma).$$

Der Beweis für die Dirichlet-Spur kann [38, Thm. 3.38] entnommen werden.

Für die Neumann-Spur muss mehr getan werden, nach [17, S. 280 ff.] kann aber auch hier ein Hilbert-Raum definiert werden, auf dem eine beschränkte Fortsetzung ∂_n der Neumann-Spur existiert.
Damit können die Übergänge vom Äußeren ins Innere beziehungsweise vom Inneren ins Äußere des Gebiets betrachtet werden. Diese Übergänge, also die Differenz der äußeren und inneren Spur, werden als *Sprünge* bezeichnet, die geltenden Beziehungen als *Sprungrelationen*. Die auf den Rand übertragenen Potentiale werden zu *Operatoren* [22, S. 40] [48, S. 4-5] [17, S. 113] .

Definition und Lemma 1.22 (Einfach- und Doppelschichtoperator)
Der Einfachschichtoperator $S_\kappa : H^{-\frac{1}{2}}(\Gamma) \to H^{\frac{1}{2}}(\Gamma)$ ist durch

$$S_\kappa := \gamma V_\kappa$$

gegeben und kann für eine Dichte w als schwach singuläres Oberflächenintegral mit

$$S_\kappa[w](x) = \int_\Gamma u_\kappa(x,y)w(y)\,\mathrm{d}y \qquad \textit{für fast alle } x \in \Gamma$$

berechnet werden. Der Doppelschichtoperator $D_\kappa : H^{\frac{1}{2}}(\Gamma) \to H^{\frac{1}{2}}(\Gamma)$ ist durch

$$D_\kappa := \gamma K_\kappa$$

gegeben und kann für eine Dichte v mit

$$D_\kappa[v](x) = \int_\Gamma v(y)\tfrac{\partial}{\partial n(y)}u_\kappa(x,y)\,\mathrm{d}y \qquad \textit{für fast alle } x \in \Gamma$$

berechnet werden. Das Integral ist dabei im Sinne des Cauchy'schen Hauptwerts[xvi] *als*

$$D_\kappa[v](x) = \lim_{\epsilon\to 0}\int_{\Gamma\setminus B_\epsilon(x)} v(y)\tfrac{\partial}{\partial n(y)}u_\kappa(x,y)\,\mathrm{d}y \qquad \textit{für fast alle } x \in \Gamma$$

zu verstehen.

Dass die beiden Operatoren nur für fast alle $x \in \Gamma$ definiert sind, liegt daran, dass Lipschitz-Gebiete betrachtet werden, für hinreichend glatte Gebiete und Dichtefunktionen existieren die Operatoren für alle $x \in \Gamma$ [17, S. 113]. Beim Anwenden der Neumann-Spur auf die Potentiale treten noch weitere Operatoren auf [17, S. 113-16][48, S. 4-5].

Definition 1.23 (Adjungierter Doppelschichtoperator)
Der adjungierte Doppelschichtoperator $D'_\kappa : H^{-\frac{1}{2}}(\Gamma) \to H^{-\frac{1}{2}}(\Gamma)$ *ist durch*

$$D'_\kappa := \partial_n V_\kappa$$

gegeben. Der Operator ist ebenfalls im Sinne des Cauchy'schen Hauptwerts zu verstehen und kann für eine Dichte w mit

$$D'_\kappa[w](x) = \int_\Gamma \tfrac{\partial}{\partial n(x)}u_\kappa(x,y)w(y)\,\mathrm{d}y \qquad \textit{für fast alle } x \in \Gamma$$

berechnet werden.

Definition 1.24 (Hypersingulärer Operator)
Der Hypersinguläre Operator $H_\kappa : H^{\frac{1}{2}}(\Gamma) \to H^{-\frac{1}{2}}(\Gamma)$ *ist durch*

$$H_\kappa := \partial_n K_\kappa$$

gegeben und kann für eine Dichte v mit

$$H_\kappa[v](x) := \tfrac{\partial}{\partial n(x)}\int_\Gamma v(y)\tfrac{\partial}{\partial n(y)}u_\kappa(x,y)\,\mathrm{d}y \qquad \textit{für fast alle } x \in \Gamma$$

[xvi] Benannt nach dem französischen Mathematiker Augustin-Louis Cauchy.

berechnet werden, er ist dabei im Sinne von

$$H_\kappa[v](x) = \lim_{y \to x, y \in \Omega} \langle \nabla K_\kappa[v], n(x) \rangle_{\mathbb{C}}$$

zu verstehen.

Der Name des adjungierten Doppelschichtoperators rührt daher, dass er unter gewissen Umständen tatsächlich dem adjungierten Operator des Doppelschichtpotentials entspricht [17, S. 114].
Mit den beiden zusätzlichen Operatoren können die folgenden *Sprungrelationen* angegeben werden [13, S. 239] [17, S. 115] [48, S. 4] .

Lemma 1.25 (Sprungrelationen Einfachschichtpotential)
Für eine Dichte w gilt

$$\gamma^{ext} V_\kappa[w](x) = \gamma^{int} V_\kappa[w](x) \qquad \textit{für fast alle } x \in \Gamma$$

sowie

$$\begin{aligned} \partial_n^{int} V_\kappa[w](x) &= \tfrac{1}{2} w(x) + D'_\kappa[w](x), \\ \partial_n^{ext} V_\kappa[w](x) &= -\tfrac{1}{2} w(x) + D'_\kappa[w](x) \end{aligned} \qquad \textit{für fast alle } x \in \Gamma.$$

Entsprechend lassen sich die Beziehungen wie folgt beschreiben

$$\begin{aligned} \gamma^{int/ext} V_\kappa = S_\kappa, \qquad \partial_n^{int/ext} V_\kappa = \pm \tfrac{1}{2} Id + D'_\kappa, \\ \partial_n^{ext} V_\kappa[w](x) - \partial_n^{int} V_\kappa[w](x) = -w(x) \qquad \textit{für fast alle } x \in \Gamma, \end{aligned} \tag{1.3.4}$$

hierbei bezeichne Id den Identitätsoperator.

Im Fall des Doppelschichtpotentials ergeben sich nachstehende Beziehungen [13, S. 239] [17, S. 115] [48, S. 4-5] .

Lemma 1.26 (Sprungrelationen Doppelschichtpotential)
Für eine Dichte v gilt

$$\begin{aligned} \gamma^{int} K_\kappa[v](x) &= -\tfrac{1}{2} v(x) + D_\kappa[v](x), \\ \gamma^{ext} K_\kappa[v](x) &= \tfrac{1}{2} v(x) + D_\kappa[v](x) \end{aligned} \qquad \textit{für fast alle } x \in \Gamma$$

sowie

$$\partial_n^{ext} K_\kappa[v](x) = \partial_n^{int} K_\kappa[v](x) \qquad \textit{für fast alle } x \in \Gamma.$$

Entsprechend lassen sich Beziehungen auch wie folgt beschreiben

$$\gamma^{int/ext} K_\kappa = \mp \tfrac{1}{2} Id + D_\kappa, \qquad \partial_n^{int/ext} K_\kappa = H_\kappa. \tag{1.3.5}$$

1.3.3 Integralformulierung und Diskretisierung

Um zu einer Integralformulierung zu gelangen, die es dann zu diskretisieren gilt, können verschiedene Ansätze gewählt werden. Bei der direkten Anwendung des Einfachschichtpotentials ergibt sich unter Verwendung der Sprungrelationen (siehe Lemma 1.25) und für eine Dichte w

$$S_\kappa[w](x) = g_D(x) \qquad \text{für fast alle } x \in \Gamma \tag{1.3.6}$$

für Dirichlet-Randwerte, während sich für Neumann-Randwerte

$$\pm\tfrac{1}{2}w(x) + D'_\kappa[w](x) = g_N(x) \qquad \text{für fast alle } x \in \Gamma$$

ergibt. Wird stattdessen direkt mit dem Doppelschichtpotential gearbeitet, ergeben sich mit dem Lemma zu den Sprungrelationen 1.26 und für eine Dichte v

$$\mp\tfrac{1}{2}v(x) + D_\kappa[v](x) = g_D(x) \qquad \text{für fast alle } x \in \Gamma$$

für Dirichlet-Randwerte beziehungsweise für Neumann-Randwerte

$$H_\kappa[v](x) = g_N(x) \qquad \text{für fast alle } x \in \Gamma.$$

In all diesen Fällen muss die Lösung anschließend mit Hilfe der berechneten Dichtefunktion rekonstruiert werden. Die *direkten Darstellungsformeln* ermöglichen es, diesen Schritt zu umgehen. Sie sind zunächst komplizierter, arbeiten dafür aber direkt mit der Lösung, über die meist mehr bekannt ist als über die Dichtefunktion bei den indirekten Ansätzen. Diese Ansätze führen zu den direkten Darstellungsformeln mit Lösungen u_1 im Außenraum Ω_a und u_2 im Innenraum Ω_i [17, S. 116 ff.][22, S. 36-40]

$$\begin{aligned} u_1(x) &= -V_\kappa[\partial_n^{ext}u_1](x) + K_\kappa[\gamma^{ext}u_1](x) && \text{für } x \in \Omega_a, \\ u_2(x) &= V_\kappa[\partial_n^{int}u_2](x) - K_\kappa[\gamma^{int}u_2](x) && \text{für } x \in \Omega_i, \end{aligned}$$

wobei jeweils nur eine der beiden Spuren der Lösung bekannt ist. In beiden Fällen werden so zwei Randintegralgleichungen erhalten, im Falle des Außenraumproblems wären dies

$$\begin{aligned} \gamma^{ext}u_1 &= -S_\kappa[\partial_n^{ext}u_1] + \left(\tfrac{1}{2}Id + D_\kappa\right)[\gamma^{ext}u_1] \\ \partial_n^{ext}u_1 &= \left(\tfrac{1}{2}Id - D'_\kappa\right)[\partial_n^{ext}u_1] + H_\kappa[\gamma^{ext}u_1]. \end{aligned}$$

Nach dem Einsetzen der bekannten Randwerte muss noch eine der beiden Gleichungen nach den unbekannten Randwerten aufgelöst werden. Aufgrund des Identitätsoperators entstehen jeweils eine Integralgleichung 1. und 2. Art[xvii].

[xvii] Bei einer Integralgleichung 1. Art ist die gesuchte Funktion nur unter dem Integral zu finden, bei einer 2. Art auch als skalares Vielfaches außerhalb.

Als Kurzschreibweise für die Integralgleichungen hat sich die *Calderón Projektion*[xviii] bewährt

$$P^{ext} = \tfrac{1}{2}Id + \widehat{P} \qquad\qquad P^{int} = \tfrac{1}{2}Id - \widehat{P},$$

wobei $\widehat{P}$ durch

$$\widehat{P} = \begin{pmatrix} D_\kappa & -S_\kappa \\ H_\kappa & -D'_\kappa \end{pmatrix}$$

gegeben ist [17, S. 117].
Direkte Ansätze für den Außenraum können versteckte Schwierigkeiten mit sich bringen. Handelt es sich bei der betrachteten Wellenzahl κ um die Wurzel eines Eigenwerts des Laplace-Operators des Innenraumproblems, sind die entsprechenden Einfach- und Doppelschichtoperatoren nicht invertierbar, obwohl das Außenraumproblem eine Lösung besitzt [13, S. 246] [17, S. 119] [22, S. 49 f.].
Da die Möglichkeit, eine Lösung beim direkten Ansatz zu finden, nicht für alle Wellenzahlen κ garantiert werden kann, liegt es nahe, eine Linearkombination von Operatoren zu nutzen. Bekannte Möglichkeiten der Linearkombinationsbildung stellen die Ansätze von Brakhage und Werner [11] sowie Burton und Miller [15] dar. Diese leiden jedoch unter gewissen Umständen an Stabilitätsproblemen, so dass stabilisierte Varianten entwickelt wurden, zum Beispiel in [22].

Angenommen, es wurde ein Ansatz gewählt, für den sichergestellt ist, dass eine Lösung existiert und bestimmt werden kann. Dann stellt sich die Frage zur konkreten Herangehensweise.
Der Weg zum Lösen der Integralgleichung wird mit Hilfe einer Variationsformulierung, einer Diskretisierung des Rands sowie anschließendem Galerkin-Ansatz[xix] beschritten. Auf diese Weise wird das Problem in ein lineares Gleichungssystem überführt, und für lineare Gleichungssysteme stehen diverse verschiedene Verfahren zum Lösen zur Verfügung.
Das Vorgehen soll hier beispielhaft und in Kürze für das Dirichlet-Problem mit direkter Anwendung des Einfachschichtpotentials (1.3.3) vorgeführt werden.
Nach (1.3.6) ergibt sich mit der Dichte w die folgende Gleichung

$$S_\kappa[w](x) = \int_\Gamma u_\kappa(x,y)w(y)\,\mathrm{d}y = g_D(x) \qquad \text{für fast alle } x \in \Gamma.$$

[xviii] Namensgeber ist der argentinische Mathematiker Alberto Pedro Calderón, der auf dem Gebiet der singulären Integrale forschte.
[xix] Eingeführt durch den sowjetischen Ingenieur und Mathematiker Boris Grigorjewitsch Galerkin.

Die Multiplikation mit einer Testfunktion $v \in H^{-\frac{1}{2}}(\Gamma)$ und anschließendes Integrieren liefert eine Variationsformulierung

$$\int_\Gamma S_\kappa[w](x)\overline{v(x)}\,\mathrm{d}x = \int_\Gamma \int_\Gamma u_\kappa(x,y)w(y)\overline{v(x)}\,\mathrm{d}y\,\mathrm{d}x = \int_\Gamma g_D(x)\overline{v(x)}\,\mathrm{d}x.$$

Die Interpretation des Doppelintegrals als Sesquilinearform $a(\cdot,\cdot) : H^{-\frac{1}{2}}(\Gamma) \times H^{-\frac{1}{2}}(\Gamma) \to \mathbb{C}$ und der rechten Seite als Funktion $f(\cdot)$ liefert eine Problemformulierung, bei der eine Funktion $w \in H^{-\frac{1}{2}}(\Gamma)$ gesucht ist, die

$$a(w,v) = f(v) \qquad \text{für alle } v \in H^{-\frac{1}{2}}(\Gamma)$$

erfüllt. Die mit dem Einfachschichtoperator gebildete Sesquilinearform ist beschränkt [17, S. 142 f.] und koerziv [48, S. 6], es existieren somit zwei Konstanten $K_s, K_k \in \mathbb{R}_{>0}$ mit

$$|a(w,v)| \leq K_s \|w\|_{H^{-\frac{1}{2}}(\Gamma)} \|v\|_{H^{-\frac{1}{2}}(\Gamma)}, \qquad |a(w,w)| \geq K_k \|w\|^2_{H^{-\frac{1}{2}}(\Gamma)}.$$

Damit liefert der Satz von *Lax-Milgram* [21, Satz 2.29] die eindeutige Lösbarkeit der Variationsformulierung.

Die Bearbeitung am Computer erfordert eine Diskretisierung des kontinuierlichen Problems, daher wird für die weitere Verarbeitung ein Galerkin-Ansatz verwendet. Mit einem Teilraum $H_n \subset H^{-\frac{1}{2}}(\Gamma)$ mit endlicher, reellwertiger Basis $\{\phi_i\}_{i\in I}$ für $I \subset \mathbb{N}$ wird die Problemstellung in ihre diskrete Formulierung

$$a(w_n,v_n) = f(v_n) \qquad \text{für alle } v_n \in H_n$$

überführt und entsprechend die diskrete Lösung $w_n \in H_n$ gesucht. Da H_n ein Teilraum von $H^{-\frac{1}{2}}(\Gamma)$ ist, bleiben die Eigenschaft der Beschränktheit und Koerzivität der Sesquilinarform $a(\cdot,\cdot)$ auch im Diskreten erhalten. Somit liefert der Satz von Lax-Milgram[xx] weiterhin die eindeutige Lösbarkeit, während das Lemma von *Céa*[xxi] [21, Gleichung 7.30] garantiert, dass die diskrete Lösung w_n die bestmögliche im Funktionsraum H_n ist. Die reellwertige Funktionsbasis $\{\phi_i\}_{i\in I}$ erlaubt es, die Funktionen v_n, w_n mit

$$v_n = \sum_{j\in I} y_j\phi_j, \qquad w_n = \sum_{i\in I} x_i\phi_i$$

auszudrücken. Das Einsetzen in die diskrete Formulierung sowie das Ausnutzen der Eigenschaften von Sesquilinearformen und linearen Funktionen führt zu

$$a(\sum_{i\in I} x_i\phi_i, \sum_{j\in I} y_j\phi_j) = \sum_{i\in I}\sum_{j\in I} a(\phi_i,\phi_j)x_i\overline{y_j}, \qquad f(\sum_{i\in I} y_i\phi_i) = \sum_{i\in I} f(\phi_i)\overline{y_i}.$$

[xx] Benannt nach dem ungarischen Mathematiker Peter David Lax und dem US-amerikanischen Mathematiker Arthur Norton Milgram, welche eine erste Version des Satzes bewiesen.

[xxi] Zu Ehren des französischen Mathematikers Jean Céa benannt.

Die Interpretation der Werte $a(\phi_i, \phi_j)$ als Einträge einer Matrix $A \in \mathbb{C}^{I \times I}$ und $f(\phi_i)$ als Einträge eines Vektors $b \in C^I$ sowie

$$x := \sum_{i \in I} x_i e_i, \qquad\qquad y := \sum_{j \in I} y_j e_j,$$

wobei e_i der i-te kanonische Einheitsvektor sei, liefert

$$\langle Ax, y \rangle_{\mathbb{C}} = \langle b, y \rangle_{\mathbb{C}} \qquad\qquad \text{für alle } y \in \mathbb{C}^I$$

und damit das lineare Gleichungssystem

$$Ax = b.$$

2 Der Ansatz

Mit der diskreten Formulierung stellt sich die Frage, wie das entstandene lineare Gleichungssystem auf effiziente Art und Weise gelöst werden kann. Den auftretenden, oftmals sehr großen, vollbesetzten Matrizen liegt jedoch eine gewisse Struktur zugrunde, die genutzt werden kann, um diese Matrizen möglichst effizient auf dem Computer zu verarbeiten. Eine Möglichkeit der Verarbeitung stellen die *hierarchischen Matrizen* dar [29] [4].
Die Kernidee ist, die zu behandelnde Matrix aufzuteilen, da einige Teile der Matrix speichergünstig approximiert werden können, während andere sich nicht ohne große Verluste in der Genauigkeit approximativ darstellen lassen. Die gut approximierbaren Teilmatrizen werden dann blockweise mit Produkten von Matrizen niedrigeren Rangs dargestellt. Wird eine Teilmatrix als Produkt von zwei Niedrigrang-Matrizen dargestellt, erhalten wir $\mathcal{H}$-Matrizen. Dabei steht das ‚$\mathcal{H}$' für die Hierarchie, die der Zerteilung in Teilmatrizen zugrundeliegt. Noch effizienter und ebenfalls etabliert ist ein Produkt aus drei Matrizen, einer sehr kleinen in der Mitte und zwei rechteckigen schmalen, wobei die äußeren rechteckigen zusätzlich, um Speicher zu sparen, einer gewissen Schachtelung unterliegen. In diesem Fall liegen zwei Hierarchien vor und wir erhalten $\mathcal{H}^2$-Matrizen.
Die Helmholtz-Gleichung im hochfrequenten Fall erfordert noch eine Modifikation der Standard-$\mathcal{H}^2$-Matrizen, die zur Spezialisierung, den $\mathcal{RH}^2$-Matrizen, führt [6] [3].

Auf Grund der Verwendung der Potentiale zusammen mit der Galerkin-Diskretisierung (siehe Kapitel 1.3.3) liegt den auftretenden Matrizen eine sogenannte *Kernfunktion*, hier die Fundamentallösung (1.1.3) selbst oder Derivate dieser, zugrunde.
Im Fall des Einfachschichtoperators entspricht die Kernfunktion g_e der Fundamentallösung u_κ der Helmholtz-Gleichung selbst

$$g_e(x,y) := \frac{e^{\mathrm{i}\kappa\|x-y\|_2}}{4\pi\|x-y\|_2} \qquad \text{für fast alle } x,y \in \Gamma. \tag{2.0.1}$$

Die Kernfunktion des Doppelschichtoperators ist durch die Normalenableitung entlang der äußeren Normalen in y

$$g_d(x,y) := \frac{\partial}{\partial n(y)} \frac{e^{\mathrm{i}\kappa\|x-y\|_2}}{4\pi\|x-y\|_2} \qquad \text{für fast alle } x,y \in \Gamma$$

gegeben. Falls der Rand Γ glatt ist, bei Lipschitz-Gebieten ist dies zumindest stückweise gegeben, kann die Normalenableitung auch mit Hilfe des Skalarprodukts und des Gradienten

∇_y in y ausgedrückt werden [33, S. 16]

$$g_d(x,y) := \langle \nabla_y g_e(x,y), n(y) \rangle_{\mathbb{C}} \qquad \text{für fast alle } x, y \in \Gamma.$$

Bestimmen des Gradienten von g_e führt zur Schreibweise

$$\begin{aligned} g_d(x,y) &= \langle x-y, n(y) \rangle_2 \tfrac{e^{\mathrm{i}\kappa\|x-y\|_2}}{4\pi\|x-y\|_2^3}(1-\mathrm{i}\kappa\|x-y\|_2) \\ &= g_e(x,y)\langle x-y, n(y)\rangle_2 \tfrac{1-\mathrm{i}\kappa\|x-y\|_2}{\|x-y\|_2^2}. \end{aligned}$$

Eine gängige und simple Methode zur Generierung der Matrixapproximation nutzt die Polynom-Interpolation. Aus diesem Grund werden zunächst die Grundlagen der Polynom-Interpolation vorgestellt. Im Anschluss daran werden die Kernfunktionen modifiziert, um die Effizienz der Polynom-Interpolation im hochfrequenten Bereich zu verbessern. Da Kernfunktionen mit Singularitäten betrachtet werden, ist eine Approximation nicht überall möglich. Deshalb werden für die modifizierten Kernfunktionen notwendige Bedingungen für die Approximation formuliert. Entsprechend müssen Teilgebiete, auf denen die Approximation möglich ist, bestimmt und die verwendeten Hierarchien aufgebaut werden, womit sich die letzten beiden Teile dieses Kapitels beschäftigen.

2.1 Interpolation

Eine Möglichkeit, eine Systemmatrix mit zugrundeliegender Kernfunktion zu approximieren, stellt die Interpolation, präziser die *Polynom-Interpolation*, dar.
Definiere den Raum der Polynome, die höchstens von Grad $m \in \mathbb{N}_0$ sind, durch

$$\Pi_m := \operatorname{span}\left\{ x \mapsto \sum_{i=0}^{m} a_i x^i \mid a_0, \ldots, a_m \in \mathbb{C} \right\}.$$

Betrachte zunächst den Fall einer Funktion f mit eindimensionalem Definitionsbereich $[a,b]$ mit $a, b \in \mathbb{R}$ und $a < b$. Zu einer vorher festgelegten Interpolationsordnung $m \in \mathbb{N}_0$, bezeichne im Folgenden M die Menge $M := \{0, \ldots, m\}$. Für paarweise verschiedene Stützstellen $x_0 < x_1 < \cdots < x_m$ auf $[a,b]$ wird ein Polynom $p \in \Pi_m$ mit

$$p(x_\mu) = f(x_\mu) \qquad \text{für alle } \mu \in M$$

gesucht. Diese Problemstellung lässt sich immer eindeutig lösen (siehe [24, Thm. 2.1.1.1]), weshalb im Folgenden von dem Interpolationspolynom gesprochen wird.
Der einfachste Weg, das Interpolationspolynom zu konstruieren, nutzt die *Lagrange-Polynome*[i].

[i] Benannt nach dem italienischen Mathematiker und Astronomen Joseph-Louis Lagrange.

Definition 2.1 (Lagrange-Polynome)
Seien zu $m \in \mathbb{N}_0$ paarweise verschiedene Stützstellen $x_0 < x_1 < \cdots < x_m$ auf $[a,b]$ gegeben. Für $\mu \in M$ ist das μ-te Lagrange-Polynom *zu diesen Stützstellen definiert durch*

$$\ell_\mu(x) = \prod_{\substack{\nu=0 \\ \nu \neq \mu}}^{m} \frac{x - x_\nu}{x_\mu - x_\nu} \qquad \textit{für alle } x \in \mathbb{R},$$

es ist somit ein reelles Polynom von Grad m [18, G. 8.6] .

Die Eigenschaft, welche die besondere Eignung der Lagrange-Polynome für diese Aufgabe darstellt, ist, dass sie in den Stützstellen jeweils die Werte 0 oder 1 annehmen. Es gilt

$$\ell_\mu(x_\nu) = \begin{cases} 1 & \text{falls } \mu = \nu, \\ 0 & \text{sonst} \end{cases} \qquad \text{für alle } \nu \in M, \mu \in M,$$

weshalb das Interpolationspolynom $p \in \Pi_m$ direkt durch

$$p(x) = \sum_{\mu \in M} f(x_\mu)\ell_\mu(x) \qquad \text{für } x \in [a,b]$$

dargestellt werden kann. Zur Verkürzung der Notation führe einen *Interpolationsoperator* zunächst jedoch nur auf dem Referenzintervall $[-1,1]$ ein.

Definition 2.2 (Eindimensionaler Interpolationsoperator Referenzintervall)
Seien eine feste Interpolationsordnung $m \in \mathbb{N}_0$ und paarweise verschiedene und aufsteigend sortierte Stützstellen $\{\hat{x}_\mu\}_{\mu \in M} \in [-1,1]$ gegeben. Für alle $\mu \in M$ sei $\hat{\ell}_\mu$ das μ-te Lagrange-Polynom zu den Stützstellen $\{\hat{x}_\mu\}_{\mu \in M}$. Der dazugehörige eindimensionale Interpolationsoperator $\mathfrak{I}_{[-1,1]}$ *ist definiert durch*

$$\mathfrak{I}_{[-1,1]} : C([-1,1]) \to \Pi_m \qquad\qquad f \mapsto \sum_{\mu \in M} f(\hat{x}_\mu)\hat{\ell}_\mu.$$

Somit lässt sich das Interpolationspolynom p auf $[-1,1]$ mit $\mathfrak{I}_{[-1,1]}[f]$ schreiben [18, Satz 8.3].

Der so definierte Interpolationsoperator ist linear und auf dem Referenzintervall $[-1,1]$ beschränkt, denn für eine Funktion $f \in C([-1,1])$ gilt für alle $\hat{x} \in [-1,1]$

$$\begin{aligned} \left|\mathfrak{I}_{[-1,1]}[f](\hat{x})\right| = \left|\sum_{\mu \in M} f(\hat{x}_\mu)\hat{\ell}_\mu(\hat{x})\right| &\leq \max\left\{|f(\hat{x})| \mid \hat{x} \in [-1,1]\right\} \sum_{\mu \in M} \left|\hat{\ell}_\mu(\hat{x})\right| \\ &\leq \|f\|_{\infty,[-1,1]} \max\left\{ \sum_{\mu \in M} \left|\hat{\ell}_\mu(\hat{x})\right| \;\middle|\; \hat{x} \in [-1,1] \right\} \end{aligned}$$

und damit

$$\left\|\mathfrak{I}_{[-1,1]}[f]\right\|_{\infty,[-1,1]} \leq \|f\|_{\infty,[-1,1]} \max\left\{\sum_{\mu\in M}\left|\hat{\ell}_\mu(\hat{x})\right| \mid \hat{x}\in[-1,1]\right\}. \tag{2.1.1}$$

Definition 2.3 (Lebesgue-Konstante)
Die Lebesgue- *oder* Stabilitätskonstante *des Interpolationsoperators* $\mathfrak{I}_{[-1,1]}$ *bezüglich der Stützstellen* $\hat{x}_0 < \hat{x}_1 < \cdots < \hat{x}_m$ *auf* $[-1,1]$ *ist gegeben durch*

$$\Lambda_m := \max\left\{\sum_{\mu\in M}\left|\hat{\ell}_\mu(\hat{x})\right| \mid \hat{x}\in[-1,1]\right\},$$

für sie gilt

$$\left\|\mathfrak{I}_{[-1,1]}\right\|_{op,C([-1,1])\leftarrow C([-1,1])} \leq \Lambda_m,$$

wobei $\|\cdot\|_{op,C([-1,1])\leftarrow C([-1,1])}$ *die Operatornorm bezeichne [19, S. 203 ff.].*

Bemerkung 6 (Stabile Interpolation): *Es gibt Interpolationsansätze für die Konstanten* $\Lambda, \lambda \in \mathbb{R}_{\geq 1}$ *existieren, so dass*

$$\Lambda_m \leq \Lambda(m+1)^\lambda \qquad \text{für alle } m\in\mathbb{N}_0 \tag{2.1.2}$$

gilt.

Die Lebesgue-Konstante hängt nur von der relativen Lage der Stützstellen zueinander ab, nicht vom Intervall selbst. Die optimale Wahl der Stützstellen, welche dann auf das jeweilige Intervall transformiert werden, ist jedoch nicht trivial. Dennoch kann die Lebesgue-Konstante mit der Wahl der *Tschebyscheff-Knoten*[ii] klein gehalten werden [19, S. 204].
Die Tschebyscheff-Knoten sind die Nullstellen der *Tschebyscheff-Polynome*, welche sich leicht über die Rekursion

$$T_n(z) = \begin{cases} 1 & \text{für } n=0, \\ z & \text{für } n=1, \\ 2zT_{n-1}(z) - T_{n-2}(z) & \text{sonst} \end{cases} \qquad \text{für alle } n\in\mathbb{N}_0, z\in\mathbb{C} \tag{2.1.3}$$

bestimmen lassen [34, S. 284]. Für alle $n\in\mathbb{N}_0$ ist das Polynom T_n von Grad n und kann auf dem Einheitsintervall über

$$T_n(x) = \cos(n\arccos(x)) \qquad \text{für alle } x\in[-1,1] \tag{2.1.4}$$

dargestellt werden [34, S. 284]. Die Nullstellen der Tschebyscheff-Polynome lassen sich auf dem Intervall $[-1,1]$ besonders leicht angeben [19, S. 204].

[ii] Die Knoten sind, ebenso wie die Polynome, nach dem russischen Mathematiker Pafnuty Lwowich Tschebyscheff benannt, der sich unter anderem mit der Interpolation beschäftigte.

Definition 2.4 (Tschebyscheff-Knoten)
Für $m \in \mathbb{N}_0$ sind die $m+1$ Tschebyscheff-Knoten *auf dem Intervall $[-1,1]$ gegeben durch*

$$\hat{x}_\mu := \cos\left(\tfrac{2(m-\mu)+1}{2(m+1)}\pi\right) \qquad \text{für alle } \mu \in M.$$

Bemerkung 7 (Lebesgue-Konstante): *Wie in [43, S. 14 ff.] gezeigt, führen die Tschebyscheff-Knoten zu einer Lebesgue-Konstante mit*

$$\Lambda_m \leq \tfrac{2}{\pi}\ln(m+1)+1.$$

Es gilt offenbar $\Lambda_m \leq m+1$, so dass in der Notation von Bemerkung 6 für $\Lambda = 1, \lambda = 1$

$$\Lambda_m \leq \Lambda(m+1)^\lambda \qquad \text{für alle } m \in \mathbb{N}_0 \tag{2.1.5}$$

folgt.

Gesucht war ein Interpolationspolynom auf dem Intervall $[a,b]$, der bisher nur auf $[-1,1]$ definierte Interpolationsoperator kann direkt auf das allgemeine Intervall $[a,b]$ übertragen werden.

Definition 2.5 (Eindimensionaler Interpolationsoperator)
Seien eine feste Interpolationsordnung $m \in \mathbb{N}_0$ und paarweise verschiedene und aufsteigend sortierte Stützstellen $\{x_\mu\}_{\mu\in M} \in [a,b]$ gegeben. Der dazugehörige Interpolationsoperator $\mathfrak{I}_{[a,b]}$ ist durch

$$\mathfrak{I}_{[a,b]} : C([a,b]) \to \Pi_m \qquad f \mapsto \sum_{\mu\in M} f(x_\mu)\ell_\mu$$

definiert.

Mit Hilfe der bijektiven Abbildung

$$\Phi_{[a,b]} : [-1,1] \to [a,b], \qquad \hat{x} \mapsto \tfrac{b+a}{2} + \tfrac{b-a}{2}\hat{x} \tag{2.1.6}$$

kann der Interpolationsoperator auf $[a,b]$ auf den Interpolationsoperator auf dem Referenzintervall zurückgeführt werden und ist damit ebenfalls linear.
Seien $\{\hat{x}_\mu\}_{\mu\in M}$ Stützstellen auf $[-1,1]$, dann ist zu jedem $\mu \in M$ die korrespondierende Stützstelle auf $[a,b]$ durch $x_\mu = \Phi_{[a,b]}(\hat{x}_\mu)$ gegeben. Für $x \in [a,b]$ sei $\hat{x} = \Phi^{-1}_{[a,b]}(x)$, dann folgt für alle $\mu \in M$, dass die Lagrange-Polynome

$$\ell_\mu(x) = \prod_{\substack{\nu=0\\ \nu\neq\mu}}^{m} \tfrac{x-x_\nu}{x_\mu-x_\nu} = \prod_{\substack{\nu=0\\ \nu\neq\mu}}^{m} \tfrac{\Phi_{[a,b]}(\hat{x})-\Phi_{[a,b]}(\hat{x}_\nu)}{\Phi_{[a,b]}(\hat{x}_\mu)-\Phi_{[a,b]}(\hat{x}_\nu)} = \prod_{\substack{\nu=0\\ \nu\neq\mu}}^{m} \tfrac{\hat{x}-\hat{x}_\nu}{\hat{x}_\mu-\hat{x}_\nu} = \hat{\ell}_\mu(\hat{x})$$

erfüllen. Entsprechend erfüllt der eindimensionale Interpolationsoperator für alle $x \in [a,b]$ mit $\hat{x} = \Phi_{[a,b]}^{-1}(x)$

$$\mathfrak{I}_{[a,b]}[f](x) = \sum_{\mu \in M} f(x_\mu)\ell_\mu(x) = \sum_{\mu \in M} f \circ \Phi_{[a,b]}(\hat{x}_\mu)\hat{\ell}_\mu(\hat{x}) = \mathfrak{I}_{[-1,1]}[f \circ \Phi_{[a,b]}](\hat{x})$$

und es gilt folgende Beziehung

$$\mathfrak{I}_{[a,b]}[f] = \mathfrak{I}_{[-1,1]}[f \circ \Phi_{[a,b]}] \circ \Phi_{[a,b]}^{-1}.$$

Damit kann die Stabilitätsaussage für transformierte Stützstellen vom Referenzintervall auf $[a,b]$ übertragen werden

$$\begin{aligned}\left\|\mathfrak{I}_{[a,b]}[f]\right\|_{\infty,[a,b]} &= \left\|\mathfrak{I}_{[-1,1]}[f \circ \Phi_{[a,b]}]\right\|_{\infty,[-1,1]} \leq \Lambda_m \left\|f \circ \Phi_{[a,b]}\right\|_{\infty,[-1,1]} \\ &= \Lambda_m \|f\|_{\infty,[a,b]}.\end{aligned}$$

Um die Kernfunktion interpolieren zu können, muss das Konzept der Interpolation auf den mehrdimensionalen Fall erweitert werden. Die Erweiterung wird mit Hilfe des Tensorprodukts durchgeführt und auf diese Weise auf den eindimensionalen Interpolationsoperator zurückgeführt [32, S. 369 f.]. Sei $m \in \mathbb{N}_0$ die Interpolationsordnung, welche für alle Koordinatenrichtungen genutzt werden soll. Verwende die Schreibweisen $\underline{d} := \{1, 2, \ldots, d\}$ und $\underline{d}_0 := \{0, 1, \ldots, d\}$ für Mengen von natürlichen Zahlen mit $d \in \mathbb{N}$. Für einen d-dimensionalen Quader $Q_d := [a_1, b_1] \times \cdots \times [a_d, b_d]$ und $i \in \underline{d}$ sei $Q_{d,i} = [a_i, b_i]$. Bezeichne mit ${}^{d,i}\mathfrak{I}_{Q_{d,i}}$ den Interpolationsoperator im d-Dimensionalen, der entlang der i-ten Koordinatenrichtung interpoliert und alle anderen Koordinatenrichtungen unberührt lässt. Die Stützstellen $\{x_{i,\mu}\}_{\mu \in M} \in Q_{d,i}$ gehen durch Transformation aus denen auf dem Intervall $[-1,1]$ hervor. Für eine Funktion $f \in C(Q_d)$ gilt dann

$${}^{d,i}\mathfrak{I}_{Q_{d,i}}[f](y) = \sum_{\mu \in M} f(y_1, \ldots, y_{i-1}, x_\mu, y_{i+1} \ldots, y_d)\ell_\mu(y_i) \quad \text{für alle } y \in Q_d.$$

Definition 2.6 (Tensorinterpolationsoperator)
Für den d-dimensionalen Quader $Q_d = [a_1, b_1] \times \cdots \times [a_d, b_d]$ definiere den Tensorinterpolationsoperator über die zugehörigen Interpolationsoperatoren der Ordnung $m \in \mathbb{N}_0$ in einer Dimension ${}^{d,i}\mathfrak{I}_{Q_{d,i}}$ für $i \in \underline{d}$ durch

$$\mathfrak{I}_{Q_d} := {}^{d,1}\mathfrak{I}_{Q_{d,1}} \circ \cdots \circ {}^{d,d}\mathfrak{I}_{Q_{d,d}}.$$

Für ein festes $m \in \mathbb{N}_0$ und einen d-dimensionalen Quader Q_d werden die Stützstellen mit Hilfe des kartesischen Produkts aus den eindimensionalen Stützstellen gebildet. Sei für $i \in \underline{d}$ die Familie der Stützstellen in der i-ten Koordinatenrichtung $\{x_{i,\mu}\}_{\mu \in M}$. Verwende

die Menge der Multiindizes $\widehat{M} := M \times \cdots \times M = M^d$. Die d-dimensionalen Stützstellen sind dann für $\widehat{\mu} \in \widehat{M}$ durch

$$\widehat{x}_{\widehat{\mu}} := (x_{1,\mu_1}, \ldots, x_{d,\mu_d})^T$$

gegeben. Entsprechend ergibt sich das Interpolationspolynom des d-dimensionalen Tensorinterpolationsoperators für eine Funktion $f \in C(Q_d)$ im Punkt $y = (y_1, \ldots, y_d)^T \in Q_d$ mit

$$\mathfrak{I}_{Q_d}[f](y) = \sum_{\widehat{\mu} \in \widehat{M}} f(\widehat{x}_{\widehat{\mu}}) \ell_{\widehat{\mu}}(y),$$

wobei das d-dimensionale Lagrange-Polynom $\ell_{\widehat{\mu}}(y)$ durch

$$\ell_{\widehat{\mu}}(y) := \ell_{1,\mu_1}(y_1) \cdots \ell_{d,\mu_d}(y_d) \tag{2.1.7}$$

gegeben ist und ℓ_{i,μ_i} für $i \in \underline{d}$ das μ_i-te Lagrange-Polynom zu den Stützstellen $\{x_{i,\mu}\}_{\mu \in M}$ bezeichne.

Die Lebesgue-Konstante kann direkt von dem eindimensionalen Interpolationsoperator auf ${}^{d,i}\mathfrak{I}_{Q_{d,i}}$ übertragen werden [32, S. 369 f.]. Sei eine Funktion mit d-dimensionalem Definitionsbereich $f \in C(Q_d)$ gegeben. Definiere für $x \in Q_d$ und $i \in \underline{d}$ eine Funktion $f_{x,i}$, die alle bis auf die i-te Komponente festhält und folglich einen eindimensionalen Definitionsbereich besitzt

$$f_{x,i}(\xi) := f(x_1, \ldots, x_{i-1}, \xi, x_{i+1}, \ldots, x_d) \qquad \text{für } \xi \in [a_i, b_i]. \tag{2.1.8}$$

Offensichtlich gilt

$${}^{d,i}\mathfrak{I}_{Q_{d,i}}[f](x) = \mathfrak{I}_{Q_{d,i}}[f_{x,i}](x_i).$$

Sei $y \in Q_d$ ein Punkt, in dem ${}^{d,i}\mathfrak{I}_{Q_{d,i}}[f]$ sein Maximum auf Q_d annimmt, dann folgt direkt nach (2.1.1) die Stabilitätsaussage für ${}^{d,i}\mathfrak{I}_{Q_{d,i}}$ mit

$$\|{}^{d,i}\mathfrak{I}_{Q_{d,i}}[f]\|_{\infty,Q_d} = \|\mathfrak{I}_{Q_{d,i}}[f_{y,i}]\|_{\infty,Q_{d,i}} \leq \Lambda_m \|f\|_{\infty,Q_d}. \tag{2.1.9}$$

Induktiv lässt sich dies auch auf den Tensorinterpolationsoperators übertragen, es gilt

$$\begin{aligned} \|\mathfrak{I}_{Q_d}[f]\|_{\infty,Q_d} &= \|{}^{d,1}\mathfrak{I}_{Q_{d,1}} \circ \cdots \circ {}^{d,d}\mathfrak{I}_{Q_{d,d}}[f]\|_{\infty,Q_d} \\ &\leq \Lambda_m \|{}^{d,2}\mathfrak{I}_{Q_{d,2}} \circ \cdots \circ {}^{d,d}\mathfrak{I}_{Q_{d,d}}[f]\|_{\infty,Q_d} \\ &\leq \Lambda_m^d \|f\|_{\infty,Q_d}. \end{aligned}$$

Ebenso lassen sich mit dem Konzept (2.1.8) die Fehler des Tensorinterpolationsoperators auf die Fehler in einer Dimension zurückführen.

Lemma 2.7
Seien ein d-dimensionaler Quader Q_d, eine Funktion $f \in C(Q_d)$ und für alle $i \in \underline{d}$ und $x \in Q_d$ die Funktion $f_{x,i} : [-1,1] \to \mathbb{C}$ gegeben mit

$$\xi \mapsto f(x_1, \ldots, x_{i-1}, \Phi_{Q_{d,i}}(\xi), x_{i+1}, \ldots, x_d).$$

Falls für den Interpolationsoperator der Ordnung $m \in \mathbb{N}_0$ ein $\epsilon \in \mathbb{R}_{>0}$ existiert mit

$$\|f_{x,i} - \mathfrak{I}_{Q_{d,i}}[f_{x,i}]\|_{\infty,[-1,1]} \leq \epsilon \qquad \text{für alle } x \in Q_d, i \in \underline{d},$$

kann der Fehler der Tensorinterpolation beschränkt werden durch

$$\|f - \mathfrak{I}_{Q_d}[f]\|_{\infty,Q_d} \leq d\Lambda_m^{d-1}\epsilon.$$

Beweis: Für alle $i \in \underline{d}$ existiert ein $y \in Q_d$, so dass die Voraussetzung

$$\left\|f - {}^{d,i}\mathfrak{I}_{Q_{d,i}}[f]\right\|_{\infty,Q_d} = \left\|f_{y,i} - \mathfrak{I}_{Q_{d,i}}[f_{y,i}]\right\|_{\infty,[-1,1]} \leq \epsilon$$

liefert. Mit Hilfe einer Teleskopsumme und unter Nutzung der Definition des Tensorinterpolationsoperators ergibt sich

$$\begin{aligned}
\|f - \mathfrak{I}_{Q_d}[f]\|_{\infty,Q_d} &= \left\|f - {}^{d,1}\mathfrak{I}_{Q_{d,1}} \circ \cdots \circ {}^{d,d}\mathfrak{I}_{Q_{d,d}}[f]\right\|_{\infty,Q_d} \\
&= \left\|\sum_{i=1}^{d} {}^{d,1}\mathfrak{I}_{Q_{d,1}} \circ \cdots \circ {}^{d,i-1}\mathfrak{I}_{Q_{d,i-1}}[f - {}^{d,i}\mathfrak{I}_{Q_{d,i}}[f]]\right\|_{\infty,Q_d} \\
&\overset{\Delta}{\leq} \sum_{i=1}^{d} \left\|{}^{d,1}\mathfrak{I}_{Q_{d,1}} \circ \cdots \circ {}^{d,i-1}\mathfrak{I}_{Q_{d,i-1}}[f - {}^{d,i}\mathfrak{I}_{Q_{d,i}}[f]]\right\|_{\infty,Q_d} \\
&\overset{(2.1.9)}{\leq} \sum_{i=1}^{d} \Lambda_m^{i-1} \left\|f - {}^{d,i}\mathfrak{I}_{Q_{d,i}}[f]\right\|_{\infty,Q_d} \\
&\leq d\Lambda_m^{d-1}\epsilon.
\end{aligned}$$

$\square$[3,L.3.3]

Damit auch die Interpolation des Doppelschichtoperators betrachtet werden kann, sind zusätzlich Aussagen zur Differenzierung von Interpolationspolynomen nötig. Der mehrdimensionale Fall kann wie zuvor auf den eindimensionalen zurückgeführt werden. Nach dem Satz von Schwarz ist es für eine hinreichend oft differenzierbare Funktion f egal, in welcher Reihenfolge die partiellen Ableitungen betrachtet werden, es gilt $\partial_i\partial_j f = \partial_j\partial_i f$. Wird die partielle Ableitung in einer anderen Richtung betrachtet, als die Funktion interpoliert, so gilt ebenfalls für alle $x \in Q_d$

$$\partial_j(f_{x,i} - \mathfrak{I}_{[-1,1]}[f_{x,i}]) = \partial_j f_{x,i} - \mathfrak{I}_{[-1,1]}[\partial_j f_{x,i}] \qquad \text{für alle } i, j \in \underline{d} \text{ mit } i \neq j.$$

Dies kann leicht durch die Linearität des partiellen Differentialoperators verifiziert werden. Entsprechend gilt für eine Funktion $f \in C^1(\mathbb{R}^3)$ auf $[a_1, b_1] \times [a_2, b_2] \times [a_3, b_3] \subset \mathbb{R}^3$ mit Stützstellen $x_{1,0} < \cdots < x_{1,m} \in [a_1, b_1]$

$$\begin{aligned} {}^{3,1}\mathfrak{I}_{[a_1,b_1]}\left[\tfrac{\partial}{\partial y} f\right](x,y,z) &= \sum_{\mu=0}^{m} \tfrac{\partial}{\partial y} f(x_{1,\mu}, y, z)\ell_\mu(x) = \tfrac{\partial}{\partial y} \sum_{\mu=0}^{m} f(x_{1,\mu}, y, z)\ell_\mu(x) \\ &= \tfrac{\partial}{\partial y}\, {}^{3,1}\mathfrak{I}_{[a_1,b_1]}[f](x,y,z). \end{aligned}$$

Für die mehrdimensionale Fehleraussage bei partiellen Ableitungen der Tensorinterpolation definiere die Funktion, die alle bis auf die i-te Komponente festhält, für alle $x \in Q_d$, für ein festes $j \in \underline{d}$ und für alle $\xi \in [a_i, b_i]$ durch

$$f_{x,i}(\xi) := \begin{cases} (\partial_j f)(x_1, \ldots, x_{i-1}, \xi, x_{i+1}, \ldots, x_d) & \text{für } i \neq j, \\ f(x_1, \ldots, x_{i-1}, \xi, x_{i+1}, \ldots, x_d) & \text{sonst.} \end{cases} \tag{2.1.10}$$

Satz 2.8 (Partielle Ableitungen des Tensorinterpolationsfehlers)
Seien für $d \in \mathbb{N}$ ein Quader Q_d, ein Multiindex $\alpha \in \mathbb{N}_0^d$ mit $|\alpha| = 1$ und eine stetig differenzierbare Funktion $f : Q_d \to \mathbb{C}$ gegeben. Sei $j \in \underline{d}$ der Index mit $\alpha_j = 1$. Falls für den Interpolationsoperator der Ordnung $m \in \mathbb{N}_0$ Schranken $\epsilon, \widehat{\epsilon} \in \mathbb{R}_{\geq 0}$ existieren, so dass die Funktion $f_{x,i}$ für $i = j$

$$\left\|(f_{x,j} \circ \Phi_{Q_{d,j}})' - (\mathfrak{I}_{[-1,1]}[f_{x,j} \circ \Phi_{Q_{d,j}}])'\right\|_{\infty,[-1,1]} \leq \widehat{\epsilon} \qquad \text{für alle } x \in Q_d$$

und für alle weiteren $i \in \underline{d}$ mit $i \neq j$

$$\left\|f_{x,i} \circ \Phi_{Q_{d,i}} - \mathfrak{I}_{[-1,1]}[f_{x,i} \circ \Phi_{Q_{d,i}}]\right\|_{\infty,[-1,1]} \leq \epsilon \qquad \text{für alle } x \in Q_d$$

erfüllen, kann der Fehler durch

$$\|\partial^\alpha (f - \mathfrak{I}_{Q_d}[f])\|_{\infty,Q_d} \leq (d-1)\Lambda_m^{d-2}\epsilon + \Lambda_m^{d-1}\left|(\Phi'_{Q_{d,j}})^{-1}\right|\widehat{\epsilon}$$

beschränkt werden.

Beweis: Ohne Beschränkung der Allgemeinheit kann angenommen werden, dass $j = d$ gilt, da die eindimensionalen Interpolationsoperatoren der Tensorinterpolation kommutieren, ist es immer möglich, das j an die letzte Stelle zu tauschen.
Die Definition der Tensorinterpolation zusammen mit der Dreiecksungleichung führt zu

$$\begin{aligned} \|\partial^\alpha (f - \mathfrak{I}_{Q_d}[f])\|_{\infty,Q_d} &= \left\|\sum_{i=1}^{d} \partial^\alpha ({}^{d,1}\mathfrak{I}_{Q_{d,1}} \circ \cdots \circ {}^{d,i-1}\mathfrak{I}_{Q_{d,i-1}}[f - {}^{d,i}\mathfrak{I}_{Q_{d,i}}[f]])\right\|_{\infty,Q_d} \\ &\overset{\Delta}{\leq} \sum_{i=1}^{d} \left\|\partial^\alpha ({}^{d,1}\mathfrak{I}_{Q_{d,1}} \circ \cdots \circ {}^{d,i-1}\mathfrak{I}_{Q_{d,i-1}}[f - {}^{d,i}\mathfrak{I}_{Q_{d,i}}[f]])\right\|_{\infty,Q_d}. \end{aligned}$$

Für alle $i \in \underline{d-1}$ gilt $\alpha_i = 0$, so dass der Differentialoperator ins Argument des Interpolationsoperators in i gezogen werden kann. Mit der Stabilitätskonstante können die eindimensionalen Interpolationsoperatoren in i nach und nach abgeschätzt werden, so dass

$$\begin{aligned}
&\sum_{i=1}^{d} \left\| \partial^\alpha ({}^{d,1}\mathfrak{I}_{Q_{d,1}} \circ \cdots \circ {}^{d,i-1}\mathfrak{I}_{Q_{d,i-1}}[f - {}^{d,i}\mathfrak{I}_{Q_{d,i}}[f]]) \right\|_{\infty,Q_d} \\
&= \sum_{i=1}^{d} \left\| {}^{d,1}\mathfrak{I}_{Q_{d,1}} \circ \cdots \circ {}^{d,i-1}\mathfrak{I}_{Q_{d,i-1}}[\partial^\alpha (f - {}^{d,i}\mathfrak{I}_{Q_{d,i}}[f])]) \right\|_{\infty,Q_d} \\
&\leq \sum_{i=1}^{d} \Lambda_m^{i-1} \left\| \partial^\alpha (f - {}^{d,i}\mathfrak{I}_{Q_{d,i}}[f]) \right\|_{\infty,Q_d}
\end{aligned}$$

entsteht. Für $i \in \underline{d-1}$ gilt $\alpha_i = 0$. Wenn $y \in Q_d$ ein Punkt ist, in dem $|\partial^\alpha (f - {}^{d,i}\mathfrak{I}_{Q_{d,i}}[f])(y)|$ sein Maximum annimmt, dann folgt direkt mit der Voraussetzung

$$\left\| \partial^\alpha (f - {}^{d,i}\mathfrak{I}_{Q_{d,i}}[f]) \right\|_{\infty,Q_d} = \left\| f_{y,i} \circ \Phi_{Q_{d,i}} - \mathfrak{I}_{[-1,1]}[f_{y,i} \circ \Phi_{Q_{d,i}}] \right\|_{\infty,[-1,1]} \leq \epsilon.$$

Im Fall $i = j$ und mit einem Punkt $y \in Q_d$, in dem $|\partial^\alpha (f - {}^{d,i}\mathfrak{I}_{Q_{d,i}}[f])(y)|$ sein Maximum annimmt, gilt mit der Kettenregel

$$\begin{aligned}
\left\| \partial^\alpha (f - {}^{d,i}\mathfrak{I}_{Q_{d,i}}[f]) \right\|_{\infty,Q_d} &= \left\| \left(f_{y,j} - \mathfrak{I}_{[-1,1]}[f_{y,j}] \right)' \circ \Phi_{Q_{d,j}} \right\|_{\infty,[-1,1]} \\
&\leq \left| (\Phi'_{Q_{d,j}})^{-1} \right| \left\| (f_{y,j} \circ \Phi_{Q_{d,j}})' - (\mathfrak{I}_{[-1,1]}[f_{y,j} \circ \Phi_{Q_{d,j}}])' \right\|_{\infty,[-1,1]}.
\end{aligned}$$

Die Norm ist nach der Voraussetzung durch $\widehat{\epsilon}$ beschränkt, so dass insgesamt

$$\begin{aligned}
\|\partial^\alpha (f - \mathfrak{I}_{Q_d}[f])\|_{\infty,Q_d} &\leq \sum_{i=1}^{d} \Lambda_m^{i-1} \left\| \partial^\alpha (f - {}^{d,i}\mathfrak{I}_{Q_{d,i}}[f]) \right\|_{\infty,Q_d} \\
&\leq \sum_{i=1}^{d-1} \Lambda_m^{i-1} \epsilon + \Lambda_m^{d-1} \left| (\Phi'_{Q_{d,j}})^{-1} \right| \widehat{\epsilon} \\
&\leq (d-1) \Lambda_m^{d-2} \epsilon + \Lambda_m^{d-1} \left| (\Phi'_{Q_{d,j}})^{-1} \right| \widehat{\epsilon}
\end{aligned}$$

folgt. □

2.2 Ebene Wellen

Wünschenswert sind Kernfunktionen, die in mindestens einer Variable *asymptotisch glatt* [32, Def. 4.2.5] sind, denn dies garantiert, dass der Fehler der Interpolation exponentiell abklingt [32, S. 63 ff.].

Definition 2.9 (Asymptotisch glatt[iii])
Bezeichne eine Funktion $f : \mathbb{R}^d \times \mathbb{R}^d$ *als asymptotisch glatt vom Grad* $s \in \mathbb{N}_0$*, falls* f *auf der Menge* $\{(x, y) \mid x, y \in \mathbb{R}^d, x \neq y\}$ *unendlich oft differenzierbar ist und Konstanten* $C_{as}, c_0, r \in \mathbb{R}_{>0}$ *existieren, so dass sich die partiellen Ableitungen durch*

$$\left|\partial_x^\nu \partial_y^\mu f(x,y)\right| \leq C_{as} \frac{c_0^{|\nu+\mu|}(\nu+\mu)!|\nu+\mu|^r}{\|x-y\|_2^{|\nu|+|\mu|+s}} \qquad \begin{cases} \text{für alle } (x,y) \text{ mit } x, y \in \mathbb{R}^d, x \neq y, \\ \text{und } \nu, \mu \in \mathbb{N}_0^d, \nu + \mu \neq 0 \end{cases}$$

beschränken lassen.

Leider erfüllen die hier betrachteten Kernfunktionen diese Eigenschaft nur ungenügend, denn die Konstante c_0 ist abhängig von κ. Das hoch oszillatorische Verhalten der Fundamentallösung u_κ für große Wellenzahlen κ erweist sich als problematisch für die Approximation [32, S. 257]. Diese Eigenschaft liegt allein im Zähler der Funktion u_κ begründet. Die Abbildung 2.1 zeigt den Realteil des Zählers der Fundamentallösung für die Wellenzahl $\kappa = 15$ und lässt die Problematik erahnen. Für steigende Wellenzahlen rücken die Wellenberge der Kugelwelle näher zusammen, was einem Anstieg der Frequenz entspricht und die Problematik verschärft.

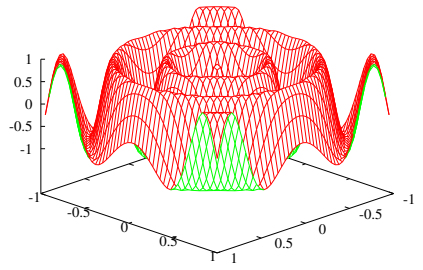


Abbildung 2.1: Realteil des Zählers der Fundamentallösung für $\kappa = 15$

Mit Verzicht auf eine von der Wellenzahl κ unabhängige exponentielle Konvergenz bietet die Interpolation zwar immer noch eine Möglichkeit der Approximation, jedoch erfordern starke Oszillationen vergleichsweise hohe Interpolationsordnungen zum Erreichen einer vorher festgelegten Genauigkeit. Der Ansatz ist folglich nicht praktikabel. Später wird sich zeigen, dass die gewählte Interpolationsordnung Einfluss auf Speicher und Rechenzeit nimmt und in Aufwandsabschätzungen teilweise mit großen Potenzen eingeht, somit sind hohe Interpolationsordnung zu vermeiden.
In dem Fall einer oszillierenden Funktion der Bauart

$$e^{\mathrm{i}\kappa\|x-y\|_2} h(x,y) \qquad \text{für } x, y \in \mathbb{R}^d,$$

[iii] Die Klassifizierung asymptotisch glatt geht auf Brandt zurück [12].

wobei h eine beliebige asymptotisch glatte Funktion ist, lässt sich das Problem, wie Brandt 1991 in [12] zeigte, durch Modifizieren der oszillierenden Funktion umgehen.
Im Kontext der Helmholtz-Gleichung sollen für die Modifikation Richtungen $c \in \mathbb{R}^3$ mit $\|c\|_2 = 1$ und dazugehörende ebene Wellen verwendet werden. Eine ebene Welle besteht aus parallel aufeinander folgenden, sich in eine feste Richtung ausbreitenden Wellenfronten. Die ebene Welle entlang der Richtung c ist durch

$$e^{\mathrm{i}\kappa\langle x-y,c\rangle_2} \qquad \text{für } x, y \in \mathbb{R}^3$$

gegeben.
Der Grundgedanke ist im Zweidimensionalen schnell erklärt. Die Wellenfronten einer Kugelwelle, die sich vom Punkte U her ausbreiten, können innerhalb eines kleinen Winkels gut durch eine ebene Welle approximiert werden. Dies ist in Abbildung 2.2 dargestellt, in der sich die ebene Welle entlang der horizontalen Achse ausbreitet, und wird durch die gestrichelten Linien angedeutet. Der zur Approximation gut geeignete, durch den Winkel aufgespannte, Bereich ist hier farblich hervorgehoben. Im Dreidimensionalen wird der für die Approximation nutzbare Bereich kegelförmig.

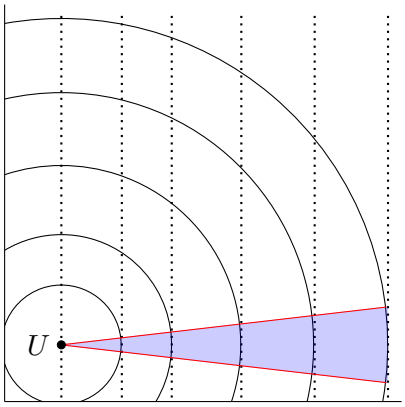


Abbildung 2.2: Approximation einer Kugelwelle mit ebenen Wellen

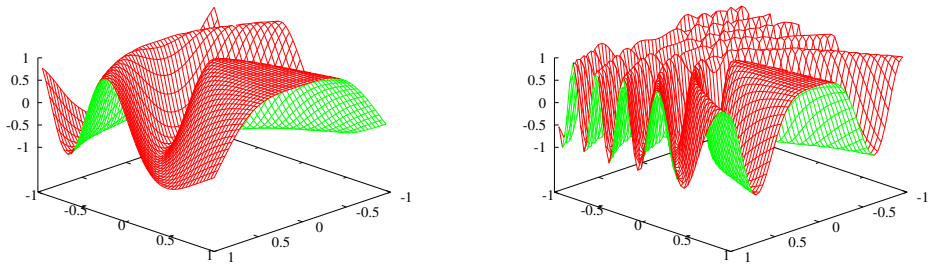
Mit genügend Ausbreitungsrichtungen für ebene Wellen lässt sich das gesamte Gebiet mit entsprechenden Kegeln abdecken, so dass eine raumfüllende Approximation mit Hilfe der ebenen Wellen möglich ist.
Sei ein $c \in \mathbb{R}^3$ mit $\|c\|_2 = 1$ gegeben. Für den Zähler der Fundamentallösung kann damit die Darstellung

$$\begin{aligned} e^{\mathrm{i}\kappa\|x-y\|_2} &= e^{\mathrm{i}\kappa\|x-y\|_2} e^{\mathrm{i}\kappa\langle x-y,c\rangle_2 - \mathrm{i}\kappa\langle x-y,c\rangle_2} \\ &= e^{\mathrm{i}\kappa(\|x-y\|_2 - \langle x-y,c\rangle_2)} e^{\mathrm{i}\kappa\langle x,c\rangle_2} \overline{e^{\mathrm{i}\kappa\langle y,c\rangle_2}} \end{aligned} \tag{2.2.1}$$

erhalten werden. Innerhalb eines Kegels um die Richtung c ist der Anteil

$$e^{\mathrm{i}\kappa(\|x-y\|_2 - \langle x-y,c\rangle_2)} \tag{2.2.2}$$

des Zählers glatt, was dazu führt, dass die durch die ebene Welle dividierte Fundamentallösung auf dem Kegel glatt ist. Die Abbildung 2.3 zeigt das Verhalten der modifizierten Funktion (2.2.2) sowie den Kegel, auf dem die Funktion glatt ist.



(a) Wellenzahl $\kappa = 5$ (b) Wellenzahl $\kappa = 15$

Abbildung 2.3: Realteil der Funktion (2.2.2) für $x - y \in [-1, 1] \times [-1, 1]$ mit $c = (1, 0)^T$

Die gewünschte Glattheit innerhalb des Kegels wird dadurch erreicht, dass der Zähler unabhängig von der Wellenzahl κ beschränkt werden kann. Um diesen Ansatz verwenden zu können, gilt es, Bedingungen zu ermitteln, unter denen die Modifikation mit einer ebenen Welle eine noch ausreichend gute Näherung darstellt.
Dazu betrachte zunächst den Exponenten ohne die imaginäre Einheit i und die Wellenzahl. Für den Rest des Arguments der modifizierten Exponentialfunktion gilt mit der Identifizierung des Skalarprodukts über den Kosinus

$$\begin{aligned}\|x-y\|_2 - \langle x-y, c\rangle_2 &= \|x-y\|_2 - \|x-y\|_2 \langle \tfrac{x-y}{\|x-y\|_2}, c\rangle_2 \\ &= \|x-y\|_2 \left(1 - \langle \tfrac{x-y}{\|x-y\|_2}, c\rangle_2\right) \\ &= \|x-y\|_2 (1 - \cos\left(\angle\left(\tfrac{x-y}{\|x-y\|_2}, c\right)\right)).\end{aligned}$$

Wichtig für einen kleinen Wert im Exponenten ist auf jeden Fall ein kleiner Winkel. Der winkelabhängige Term kann durch eine handlichere Formulierung abgeschätzt werden. Da nur kleine Winkel in Betracht gezogen werden und der Kosinus auf $\left(-\frac{\pi}{2}, \frac{\pi}{2}\right)$ positiv und maximal 1 ist, folgt

$$\cos^2(\alpha) \leq \cos(\alpha) \qquad \text{für } \alpha \in (-\tfrac{\pi}{2}, \tfrac{\pi}{2}).$$

Dies erlaubt die Abschätzung nach oben

$$1 - \cos\left(\angle\left(\tfrac{x-y}{\|x-y\|_2}, c\right)\right) \leq 1 - \cos^2\left(\angle\left(\tfrac{x-y}{\|x-y\|_2}, c\right)\right).$$

Das Anwenden der Additionstheoreme für Sinus und Kosinus zusammen mit

$$\sin^2(\alpha) \leq |\sin(\alpha)| \qquad \text{für } \alpha \in \left(-\tfrac{\pi}{2}, \tfrac{\pi}{2}\right)$$

liefern

$$1 - \cos^2\left(\angle\left(\tfrac{x-y}{\|x-y\|_2}, c\right)\right) = \sin^2\left(\angle\left(\tfrac{x-y}{\|x-y\|_2}, c\right)\right) \leq \left|\sin\left(\angle\left(\tfrac{x-y}{\|x-y\|_2}, c\right)\right)\right|.$$

An dieser Stelle bietet es sich an, das Problem geometrisch zu betrachten. Die beiden Vektoren c und $b := \frac{(x-y)}{\|x-y\|_2}$ bilden ein Dreieck, dessen dritte Seite mit $a := b - c$ bezeichnet werden soll. Mit Hilfe der Höhe h_a der Seite a und der Fläche A des Dreiecks kann der

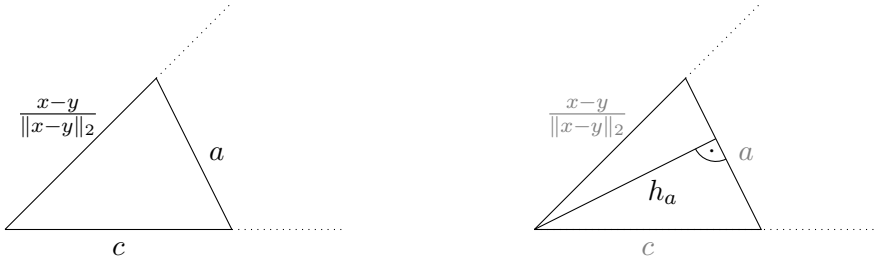


Abbildung 2.4: Skizze des gebildeten Dreiecks

Sinussatz genutzt werden

$$\frac{\|a\|_2}{\sin(\angle(b,c))} = \frac{\|a\|_2\|b\|_2\|c\|_2}{2A} = \frac{\|b\|_2\|c\|_2}{\|h_a\|_2} \qquad \Longleftrightarrow \qquad \sin\left(\angle\left(b,c\right)\right) = \frac{\|h_a\|_2\|a\|_2}{\|b\|_2\|c\|_2}.$$

Dies führt zu

$$\left|\sin\left(\angle\left(\tfrac{x-y}{\|x-y\|_2}, c\right)\right)\right| = \frac{\|h_a\|_2\|a\|_2}{\|b\|_2\|c\|_2} = \|h_a\|_2\|a\|_2.$$

Im rechtwinkligen Teildreieck, das durch die Höhe h_a und die Seite c gebildet wird, ist c die Hypotenuse, weswegen $\|h_a\|_2 \leq 1$ gilt, entsprechend folgt

$$\left|\sin\left(\angle\left(\tfrac{x-y}{\|x-y\|_2}, c\right)\right)\right| \leq \|a\|_2 = \left\|\tfrac{(x-y)}{\|x-y\|_2} - c\right\|_2.$$

Insgesamt heißt das für den Exponentialterm, dass als vorläufige Bedingung

$$\kappa\|x-y\|_2 \left\|\tfrac{(x-y)}{\|x-y\|_2} - c\right\|_2 \leq \hat{\eta}_0 \tag{2.2.3}$$

für ein geeignetes $\hat{\eta}_0 \in \mathbb{R}_{>0}$ erfüllt sein muss. Jedoch sollen die Punkte x, y innerhalb des Kegels variiert werden können, was bedeutet, dass $\kappa\|x-y\|_2$ sehr groß werden kann. Um den Zähler auch dann noch unabhängig von κ beschränken zu können, ist eine weitere Bedingung und eine tiefergehende Betrachtung nötig.
Dies lässt sich mit Hilfe einer Taylor-Entwicklung[iv] des Exponenten nachvollziehen [39, S.

[iv] Nach dem britischen Mathematiker Brook Taylor benannt.

1177]. Dazu halte die Punkte x, y fest und betrachte Punkte $\widehat{x}, \widehat{y}$ aus kleinen Umgebungen von x, y. Sei $r := \widehat{x} - x - (\widehat{y} - y)$ der Differenzvektor der Abweichung von x und $\widehat{x}$ sowie y und $\hat{y}$. Bezeichne durch u die ursprünglich betrachtete Differenz $x - y$ und schreibe damit $\widehat{x} - \widehat{y}$ auf etwas komplizierte Weise mit

$$r + u = \widehat{x} - x - (\widehat{y} - y) + x - y = \widehat{x} - \widehat{y}.$$

Führe die Taylor-Entwicklung des Exponenten der modifizierten Kernfunktion mit Richtung c zunächst noch ohne Wellenzahl und imaginäre Einheit durch und schreibe dafür den Exponenten mit Hilfe von r und u [39, G. 8]. Betrachte dann die Entwicklung von $\|r + u\|_2 - \langle r + u, c\rangle_2$ in 0 um r, es ergibt sich

$$\begin{aligned}&\|r + u\|_2 - \langle r + u, c\rangle_2 \\ &\qquad = \|u\|_2 - \langle u, c\rangle_2 + \langle r, \tfrac{u}{\|u\|_2} - c\rangle_2 + \tfrac{1}{2}\left(\tfrac{\|r\|_2^2}{\|u\|_2} - \tfrac{\langle r,u\rangle_2^2}{\|u\|_2^3}\right) + \mathcal{O}(\|r\|_2^3).\end{aligned}$$

Solange sichergestellt ist, dass $\|r\|_2$ klein genug ist, kann das Restglied $\mathcal{O}(\|r\|_2^3)$ in den Betrachtungen ignoriert werden und es reicht es aus, die anderen Terme zu untersuchen.
Der Term $\|u\|_2 - \langle u, c\rangle_2$ hängt nicht von r ab und kann damit durch eine passende Wahl von c unabhängig von κ beschränkt werden. Für die optimale Richtung c, also $c = \frac{u}{\|u\|_2}$, entfällt der Term sogar.
Der zweite Term kann für ein geeignetes η_0 durch die Bedingung

$$\kappa\|r\|_2 \left\| \tfrac{u}{\|u\|_2} - c \right\|_2 \leq \eta_0 \tag{2.2.4}$$

mit Hilfe der Cauchy-Schwarz-Ungleichung[v] beschränkt werden

$$\kappa\left|\langle r, \tfrac{u}{\|u\|_2} - c\rangle_2\right| \overset{C.S}{\leq} \kappa\|r\|_2 \left\| \tfrac{u}{\|u\|_2} - c \right\|_2.$$

Diese Bedingung entspricht für $\widehat{x} = 0$ und $\widehat{y} = 0$, der aus der Beschränkung des Winkels stammenden (2.2.3) .
Der dritte Term kann mit Hilfe der Identifikation des Skalarprodukts über der Kosinus weiter umgeformt werden

$$\begin{aligned}\tfrac{\|r\|_2^2}{\|u\|_2} - \tfrac{\langle r,u\rangle_2^2}{\|u\|_2^3} &= \tfrac{\|r\|_2^2}{\|u\|_2} - \tfrac{\|r\|_2^2\|u\|_2^2\cos^2(\angle(r,u))}{\|u\|_2^3} \\ &= \tfrac{\|r\|_2^2}{\|u\|_2}\left(1 - \cos^2(\angle(r,u)\right) \\ &= \tfrac{\|r\|_2^2}{\|u\|_2}\left(\sin^2(\angle(r,u)\right).\end{aligned}$$

Da das Quadrat des Sinus grundsätzlich in $[0, 1]$ liegt, kann der zweite Term weiter nach oben beschränkt werden, so dass eine Abschätzung von

$$\tfrac{\|r\|_2^2}{\|u\|_2} - \tfrac{\langle r,u\rangle_2^2}{\|u\|_2^3} \leq \tfrac{\|r\|_2^2}{\|u\|_2}$$

[v] Namensgebend sind die Mathematiker Augustin-Louis Cauchy und Hermann Amandus Schwarz, die verschiedene Formulierungen der Ungleichung einführten.

folgt. Entsprechend muss sichergestellt sein, dass ebenso

$$\kappa \frac{\|r\|_2^2}{\|u\|_2} \leq \eta_0 \tag{2.2.5}$$

für ein geeignetes $\eta_0 \in \mathbb{R}_{>0}$ gilt.
Unter diesen Bedingungen kann der Zähler unabhängig von κ beschränkt werden. Die bisherigen Bedingungen müssen jedoch noch weiter überarbeitet werden. Denn die Punkte $\widehat{x}, \widehat{y}$ werden später aus rechteckigen Gebieten und nicht aus kleinen Umgebungen kommen, für jedes dieser Gebiete soll nur eine Richtung verwendet werden und die Norm $\|x - y\|_2$ tritt auch im Nenner der Kernfunktion auf. Die genaue Wahl der angepassten Bedingungen für die Approximation ist Thema das Kapitels zur Zulässigkeitsbedingung 2.3.

Unter Verwendung einer ebenen Welle kann damit eine modifizierte Kernfunktion, welche innerhalb eines Kegels glatt ist, durch

$$g_{ec}(x, y) := \frac{e^{\mathrm{i}\kappa\|x-y\|_2 - \mathrm{i}\kappa\langle x-y, c\rangle_2}}{4\pi\|x-y\|_2} \tag{2.2.6}$$

definiert werden, so dass die Kernfunktion des Einfachschichtoperators mit Hilfe des Korrekturterms durch

$$g_e(x, y) = g_{ec}(x, y) e^{\mathrm{i}\kappa\langle x,c\rangle_2} \overline{e^{\mathrm{i}\kappa\langle y,c\rangle_2}} \tag{2.2.7}$$

gegeben ist. Der Korrekturterm

$$e^{\mathrm{i}\kappa\langle x,c\rangle_2} \overline{e^{\mathrm{i}\kappa\langle y,c\rangle_2}}$$

ist nach Variablen separiert, was, wie sich zeigen wird, für die Interpolation von Vorteil ist. Die Idee, die auf diese Weise modifizierte Funktion mit der Tschebyscheff-Interpolation zu approximieren, stammt von Messner, Schanz und Darve [39]. Bebendorf [2] nutzte eine Abwandlung der hierarchischen Matrizen, die im hochfrequenten Fall auf die ebene Welle als Hilfsmittel zurückgreifen, um eine ACA-basierte Approximation zu entwickeln. Der Ansatz der richtungsabhängigen hierarchischen Matrix wurde von Börm [6] vom ACA-Ansatz gelöst und zusammen mit Melenk entstanden Fehlerabschätzungen für die Approximation via Interpolation [3].

2.3 Zulässigkeitsbedingung

Zur Entscheidung, ob eine Approximation sinnvoll ist oder nicht, wird die Zulässigkeitsbedingung verwendet. Im folgenden Abschnitt geht es um die Frage, wie diese Zulässigkeitsbedingung im Fall der Helmholtz-Gleichung konkret aussieht. Schon im vorherigen Kapitel zeigte sich, dass eine Zulässigkeitsbedingung allein nicht ausreichend sein wird, insgesamt

werden drei Zulässigkeitsbedingungen notwendig werden.

Die Zulässigkeitsbedingungen werden an Eigenschaften der Gebiete geknüpft, aus denen die Punkte für die Approximation stammen. Um das Arbeiten mit den Gebieten zu erleichtern, werden ausschließlich achsenparallele Quader Q_t und Q_s betrachtet.
Der Durchmesser eines Gebiets entspricht dem maximalen Abstand, den zwei Punkte aus dem Gebiet zueinander haben können, also

$$\operatorname{diam}(Q_t) := \max\{\|x-y\|_2 \mid x,y \in Q_t\}.$$

Im Eindimensionalen ist der Durchmesser des Quaders $Q_t = [a_t, b_t]$ durch die Intervalllänge und damit durch

$$\operatorname{diam}([a_t, b_t]) = b_t - a_t$$

gegeben. Verallgemeinerungen für den zwei- und dreidimensionalen Fall lassen sich geometrisch herleiten. Eine Aussage für beliebige Dimensionen liefert das folgende Lemma.

Lemma 2.10 (Durchmesser)
Seien ein $d \in \mathbb{N}$ und ein d-dimensionaler Quader $Q_t := [a_{t,1}, b_{t,1}] \times \cdots \times [a_{t,d}, b_{t,d}]$ gegeben, dann gilt für den Durchmesser des Quaders

$$\operatorname{diam}(Q_t) = \left(\sum_{i=1}^{d} \operatorname{diam}^2([a_{t,i}, b_{t,i}])\right)^{\frac{1}{2}}.$$

Beweis: Seien ein Quader Q_t und zwei beliebige Punkte $x, y \in Q_t$ gegeben, dann gilt

$$\|x-y\|_2 = \left(\sum_{i=1}^{d} (x_i - y_i)^2\right)^{\frac{1}{2}} \leq \left(\sum_{i=1}^{d} \operatorname{diam}^2([a_{t,i}, b_{t,i}])\right)^{\frac{1}{2}}.$$

Da die beiden Punkte beliebig waren, folgt

$$\operatorname{diam}(Q_t) = \max\{\|x-y\|_2 \mid x,y \in Q_t\} \leq \left(\sum_{i=1}^{d} \operatorname{diam}^2([a_{t,i}, b_{t,i}])\right)^{\frac{1}{2}}.$$

Werden die Punkte $x, y \in Q_t$ so gewählt, dass

$$x_i - y_i = b_{t,i} - a_{t,i} = \operatorname{diam}([a_{t,i}, b_{t,i}]) \qquad \text{für alle } i \in \underline{d}$$

gilt, folgt

$$\left(\sum_{i=1}^{d} \operatorname{diam}^2([a_{t,i}, b_{t,i}])\right)^{\frac{1}{2}} = \left(\sum_{i=1}^{d} \|x_i - y_i\|_2^2\right)^{\frac{1}{2}} = \|x-y\|_2 \leq \operatorname{diam}(Q_t)$$

und somit die Behauptung. □[7]

Die Distanz zweier Quader Q_t und Q_s ist durch den minimalen Abstand zweier Punkte $x \in Q_t$ und $y \in Q_s$ definiert

$$\operatorname{dist}(Q_t, Q_s) := \min \{ \|x - y\|_2 \mid x \in Q_t, y \in Q_s \}.$$

Lemma 2.11 (Distanz)
Seien ein $d \in \mathbb{N}$ und d-dimensionale Quader $Q_t := [a_{t,1}, b_{t,1}] \times \cdots \times [a_{t,d}, b_{t,d}]$ sowie $Q_s := [a_{s,1}, b_{s,1}] \times \cdots \times [a_{s,d}, b_{s,d}]$ gegeben, dann gilt für die Distanz der Quader

$$\operatorname{dist}(Q_t, Q_s) = \left(\sum_{i=1}^{d} \operatorname{dist}^2([a_{t,i}, b_{t,i}], [a_{s,i}, b_{s,i}]) \right)^{\frac{1}{2}}.$$

Beweis: Der Beweis verläuft auf die gleiche Weise wie für den Durchmesser. □[7]

Die grundlegende Idee der Approximation ist die Trennung der Variablen x und y der Kernfunktion mit Hilfe einer Näherung durch separate Funktionen in x beziehungsweise y. Nur in wenigen Fällen besitzt die Funktion eine sogenannte *separable Entwicklung*, so dass durch die Trennung der Variablen kein Fehler entsteht. Umso wichtiger ist es, den entstehenden Fehler zu begrenzen, wobei exponentielles Abklingen bei steigender Distanz der betrachteten Punkte wünschenswert ist.
Im Fall der Helmholtz-Gleichung beinhaltet der Nenner der Kernfunktionen den Abstand der Punkte, was zu Singularitäten führt

$$\tfrac{1}{4\pi\|x-y\|_2} \to \infty \qquad \text{für} \quad x \to y.$$

Mit Konvergenz in der Nähe der Singularität kann folglich nicht gerechnet werden. Auch eine Analyse des Fehlers zeigt, dass die Gebiete, aus denen die Interpolationspunkte stammen, separiert sein und einen gewissen Abstand aufweisen müssen. Diese Bedingung findet auch in der Theorie der $\mathcal{H}^2$-Matrizen Anwendung und läuft unter der Bezeichnung der *Standardzulässigkeitsbedingung* [4, S. 46] [3]

$$\max \{\operatorname{diam}(Q_t), \operatorname{diam}(Q_s)\} \leq \eta_2 \operatorname{dist}(Q_t, Q_s) \qquad \text{für } \eta_2 \in \mathbb{R}_{>0}.$$

Diese allein reicht jedoch nicht aus.
Die Variation der Vektoren $x - y$ und $\widehat{x} - \widehat{y}$ für $x, \widehat{x} \in Q_t$ und $y, \widehat{y} \in Q_s$ darf nicht zu groß werden, damit die Richtung c für alle Kombinationen eine akzeptable Näherung der tatsächlichen Richtung darstellt. Unabhängig von der Richtung selbst gilt es dazu, (2.2.5), also

$$\kappa \tfrac{\|\widehat{x}-x-(\widehat{y}-y)\|_2^2}{\|x-y\|_2} \leq \eta_0,$$

zu erfüllen. Nehme dazu an, dass x der Mittelpunkt von Q_t ist, also $x = m_t$, und ebenso, dass y der Mittelpunkt von Q_s ist, also $y = m_s$, der Punkt $\widehat{x}$ sei ein beliebiger Punkt aus Q_t und $\widehat{y}$ ein beliebiger Punkte aus Q_s. Dann wird (2.2.5) zu

$$\kappa \frac{\|\widehat{x}-m_t-(\widehat{y}-m_s)\|_2^2}{\|m_t-m_s\|_2} \leq \eta_0.$$

Um eine gut berechenbare Schranke zu finden, forme die linke Seite der Ungleichung mit der Dreiecksungleichung weiter um

$$\kappa \frac{\|\widehat{x}-m_t-(\widehat{y}-m_s)\|_2^2}{\|m_t-m_s\|_2} \overset{\Delta}{\leq} \kappa \frac{(\|\widehat{x}-m_t\|_2+\|\widehat{y}-m_s\|_2)^2}{\|m_t-m_s\|_2}.$$

Die Normen im Zähler können gegen die halben Durchmesser abgeschätzt werden, der Nenner ist von unten durch die Distanz beschränkt

$$\kappa \frac{(\|\widehat{x}-m_t\|_2+\|\widehat{y}-m_s\|_2)^2}{\|m_t-m_s\|_2} \leq \kappa \frac{\left(\frac{1}{2}\operatorname{diam}(Q_t)+\frac{1}{2}\operatorname{diam}(Q_s)\right)^2}{\operatorname{dist}(Q_t,Q_s)} \leq \kappa \frac{\max\left\{\operatorname{diam}^2(Q_t),\operatorname{diam}^2(Q_s)\right\}}{\operatorname{dist}(Q_t,Q_s)}.$$

Dies wird ebenfalls mit dem Parameter $\eta_2 \in \mathbb{R}_{>0}$, zu der *parabolische Zulässigkeitsbedingung* umgeformt [3]

$$\kappa \max\left\{\operatorname{diam}^2(Q_t), \operatorname{diam}^2(Q_s)\right\} \leq \eta_2 \operatorname{dist}(Q_t, Q_s).$$

Zudem muss sichergestellt sein, dass eine adäquate Richtung vorhanden ist, so dass die tatsächliche Richtung in einer gewissen Genauigkeit dargestellt werden kann.
Hierfür wird die folgende Hilfsaussage über projizierte Vektoren nötig.

Lemma 2.12 (Projektion)
Seien $a, b \in \mathbb{R}^3$ mit $\|a\|_2, \|b\|_2 \geq 1$ gegeben, dann gilt

$$\left\| \frac{a}{\|a\|_2} - \frac{b}{\|b\|_2} \right\|_2 \leq \|a-b\|_2.$$

Beweis: Nehme ohne Beschränkung der Allgemeinheit an, dass $\|b\|_2 \leq \|a\|_2$ gilt. Um die Schranke für die Norm beweisen zu können, verwende den Kosinussatz für Dreiecke. Sowohl die Vektoren $a, b, (a-b)$ als auch $a\|a\|_2^{-1}, b\|b\|_2^{-1}, (a\|a\|_2^{-1} - b\|b\|_2^{-1})$ bilden je ein Dreieck, wobei der Winkel zwischen a und b sowie $a\|a\|_2^{-1}$ und $b\|b\|_2^{-1}$ identisch ist. Für den Winkel zwischen a und b ergibt sich mit dem Kosinussatz die Darstellung

$$\cos(\angle(a,b)) = \frac{\|a\|_2^2+\|b\|_2^2-\|a-b\|_2^2}{2\|a\|_2\|b\|_2},$$

während sich im Dreieck mit den normierten Vektoren folgendes ergibt

$$\begin{aligned}\cos(\angle(a,b)) &= \frac{\left\|\frac{a}{\|a\|_2}\right\|_2^2+\left\|\frac{b}{\|b\|_2}\right\|_2^2-\left\|\frac{a}{\|a\|_2}-\frac{b}{\|b\|_2}\right\|_2^2}{2\left\|\frac{a}{\|a\|_2}\right\|_2\left\|\frac{b}{\|b\|_2}\right\|_2} \\ &= \frac{2-\left\|\frac{a}{\|a\|_2}-\frac{b}{\|b\|_2}\right\|_2^2}{2}.\end{aligned}$$

Gleichsetzen der beiden Darstellungen liefert

$$\frac{\|a\|_2^2+\|b\|_2^2-\|a-b\|_2^2}{2\|a\|_2\|b\|_2} = \frac{2-\left\|\frac{a}{\|a\|_2}-\frac{b}{\|b\|_2}\right\|_2^2}{2} \iff$$
$$\left\|\frac{a}{\|a\|_2}-\frac{b}{\|b\|_2}\right\|_2^2 = 2-\frac{\|a\|_2^2+\|b\|_2^2-\|a-b\|_2^2}{\|a\|_2\|b\|_2}.$$

Dies kann weiter umgeformt werden

$$\left\|\frac{a}{\|a\|_2}-\frac{b}{\|b\|_2}\right\|_2^2 = \frac{2\|a\|_2\|b\|_2-\|a\|_2^2-\|b\|_2^2+\|a-b\|_2^2}{\|a\|_2\|b\|_2}$$

und mit Hilfe der zweiten binomischen Formeln ergibt sich

$$\left\|\frac{a}{\|a\|_2}-\frac{b}{\|b\|_2}\right\|_2^2 = \frac{\|a-b\|_2^2-(\|a\|_2-\|b\|_2)^2}{\|a\|_2\|b\|_2}.$$

Mit der Annahme, dass $\|b\|_2 \leq \|a\|_2$ erfüllt ist, kann der Zähler nach oben gegen $\|a-b\|_2^2$ abgeschätzt werden und da die Normen im Nenner nach der Voraussetzung größer als eins sind, folgt

$$\left\|\frac{a}{\|a\|_2}-\frac{b}{\|b\|_2}\right\|_2^2 = \frac{\|a-b\|_2^2-(\|a\|_2-\|b\|_2)^2}{\|a\|_2\|b\|_2} \leq \frac{\|a-b\|_2^2}{\|a\|_2\|b\|_2}$$
$$\leq \|a-b\|_2^2.$$

Die Behauptung ergibt sich durch Ziehen der Wurzel. □

Damit kann auch (2.2.4), also

$$\kappa\|\widehat{x}-x-(\widehat{y}-y)\|_2 \left\|\frac{x-y}{\|x-y\|_2}-c\right\|_2 \leq \eta_0,$$

weiter verallgemeinert werden. Zunächst ersetze x, y erneut durch die jeweiligen Mittelpunkte m_t, m_s

$$\kappa\|\widehat{x}-m_t-(\widehat{y}-m_s)\|_2 \left\|\frac{m_t-m_s}{\|m_t-m_s\|_2}-c\right\|_2 \leq \eta_0.$$

Schätze die erste Norm mit Hilfe der Dreiecksungleichung gegen die Summe der halben Durchmesser ab, so dass

$$\kappa\|\widehat{x}-m_t-(\widehat{y}-m_s)\|_2 \left\|\frac{m_t-m_s}{\|m_t-m_s\|_2}-c\right\|_2$$
$$\leq \kappa\max\{\operatorname{diam}(Q_t),\operatorname{diam}(Q_s)\} \left\|\frac{m_t-m_s}{\|m_t-m_s\|_2}-c\right\|_2$$

folgt. Verwende ein $\eta_1 \in \mathbb{R}_{>0}$ für die Bedingung der Richtungen [3]

$$\kappa\left\|\frac{m_t-m_s}{\|m_t-m_s\|_2}-c\right\|_2 \leq \frac{\eta_1}{\max\{\operatorname{diam}(Q_t),\operatorname{diam}(Q_s)\}}.$$

Dass es für die Beschränkung des Winkels vollkommen ausreichend ist, die Mittelpunkte zu verwenden, zeigt die nachstehende Betrachtung (Vergleich siehe [3, Lem. 3.9]).
Seien $x \in Q_t$ und $y \in Q_s$ beliebige Punkte. Mit der Dreiecksungleichung gilt

$$\begin{aligned}\kappa \left\| \tfrac{x-y}{\|x-y\|_2} - c \right\|_2 &\overset{\Delta}{\leq} \kappa \left\| \tfrac{x-y}{\|x-y\|_2} - \tfrac{m_t-m_s}{\|m_t-m_s\|_2} \right\|_2 + \kappa \left\| \tfrac{m_t-m_s}{\|m_t-m_s\|_2} - c \right\|_2 \\ &\leq \kappa \left\| \tfrac{x-y}{\|x-y\|_2} - \tfrac{m_t-m_s}{\|m_t-m_s\|_2} \right\|_2 + \tfrac{\eta_1}{\max\{\operatorname{diam}(Q_t),\operatorname{diam}(Q_s)\}}.\end{aligned}$$

Zunächst halte fest, dass mit der parabolischen Zulässigkeitsbedingung

$$\begin{aligned}\|x-y\|_2 \geq \operatorname{dist}(Q_t,Q_s) \geq \tfrac{\kappa \max\left\{\operatorname{diam}^2(Q_t),\operatorname{diam}^2(Q_s)\right\}}{\eta_2} \\ \|m_t-m_s\|_2 \geq \operatorname{dist}(Q_t,Q_s) \geq \tfrac{\kappa \max\left\{\operatorname{diam}^2(Q_t),\operatorname{diam}^2(Q_s)\right\}}{\eta_2}\end{aligned}$$

gilt, entsprechend folgt

$$\begin{aligned}\left\| (x-y) \tfrac{\eta_2}{\kappa \max\left\{\operatorname{diam}^2(Q_t),\operatorname{diam}^2(Q_s)\right\}} \right\|_2 \geq 1 \\ \left\| (m_t-m_s) \tfrac{\eta_2}{\kappa \max\left\{\operatorname{diam}^2(Q_t),\operatorname{diam}^2(Q_s)\right\}} \right\|_2 \geq 1.\end{aligned}$$

Somit kann das Lemma 2.12 genutzt werden, um

$$\kappa \left\| \tfrac{x-y}{\|x-y\|_2} - \tfrac{m_t-m_s}{\|m_t-m_s\|_2} \right\|_2 \leq \|(x-y)-(m_t-m_s)\|_2 \left| \tfrac{\eta_2}{\max\left\{\operatorname{diam}^2(Q_t),\operatorname{diam}^2(Q_s)\right\}} \right|$$

zu erhalten. Da alle Faktoren im Betrag positiv sind, können die Betragsstriche auch weggelassen werden. Da sowohl x als auch m_t aus Q_t stammen, liefert Umformen der Norm

$$\begin{aligned}\|(x-y)-(m_t-m_s)\|_2 &= \|(x-m_t)+(m_s-y)\|_2 \\ &\leq \max\left\{\operatorname{diam}(Q_t),\operatorname{diam}(Q_s)\right\}.\end{aligned}$$

Insgesamt ist der Fehler durch

$$\begin{aligned}&\kappa \left\| \tfrac{x-y}{\|x-y\|_2} - c \right\|_2 \\ &\leq \tfrac{\eta_2 \max\{\operatorname{diam}(Q_t),\operatorname{diam}(Q_s)\}}{\max\left\{\operatorname{diam}^2(Q_t),\operatorname{diam}^2(Q_s)\right\}} + \tfrac{\eta_1}{\max\{\operatorname{diam}(Q_t),\operatorname{diam}(Q_s)\}} \\ &= \tfrac{\eta_1+\eta_2}{\max\{\operatorname{diam}(Q_t),\operatorname{diam}(Q_s)\}}\end{aligned} \tag{2.3.1}$$

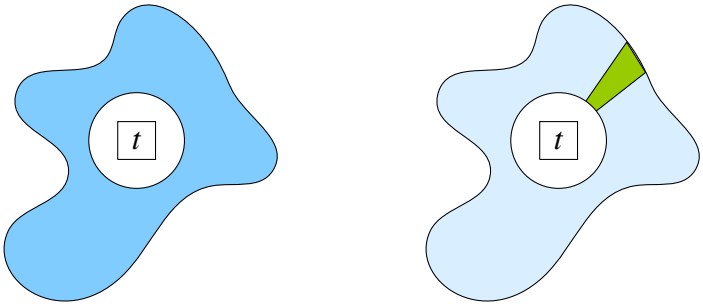
beschränkt, es reicht also vollkommen aus, eine Forderung für die Richtungen nur an die Mittelpunkte der Gebiete zu stellen.

Diese drei Anforderungen zusammen ergeben die zu erfüllenden Zulässigkeitsbedingungen. Für frei wählbare $\eta_1, \eta_2 \in \mathbb{R}_{>0}$ sind

$$\kappa \left\| \tfrac{m_t-m_s}{\|m_t-m_s\|_2} - c \right\|_2 \leq \tfrac{\eta_1}{\max\{\operatorname{diam}(Q_t),\operatorname{diam}(Q_s)\}} \tag{2.3.2a}$$

$$\kappa \max\left\{\operatorname{diam}^2(Q_t),\operatorname{diam}^2(Q_s)\right\} \leq \eta_2 \operatorname{dist}(Q_t,Q_s) \tag{2.3.2b}$$

$$\max\left\{\operatorname{diam}(Q_t),\operatorname{diam}(Q_s)\right\} \leq \eta_2 \operatorname{dist}(Q_t,Q_s) \tag{2.3.2c}$$



(a) Ohne Richtungsabhängigkeit (b) Mit Richtungsabhängigkeit

Abbildung 2.5: Vergleich der Standardzulässigkeitsbedingung mit und ohne Richtungen

zu gewährleisten [3].
Zwischen der Standardzulässigkeitsbedingung und der parabolischen ergibt sich außerdem ein Zusammenhang. Im Fall $\kappa \max\{\operatorname{diam}(Q_t), \operatorname{diam}(Q_s)\} \leq 1$ folgt aus (2.3.2c) die Bedingung (2.3.2b), da

$$\begin{aligned}\max\{\operatorname{diam}(Q_t), \operatorname{diam}(Q_s)\}\, \kappa \max\{\operatorname{diam}(Q_t), \operatorname{diam}(Q_s)\} \\ \leq \max\{\operatorname{diam}(Q_t), \operatorname{diam}(Q_s)\} \leq \eta_2 \operatorname{dist}(Q_t, Q_s)\end{aligned}$$

gilt. Falls $\max\{\operatorname{diam}(Q_t), \operatorname{diam}(Q_s)\}\, \kappa > 1$ kann entsprechend zurück umgeformt werden und (2.3.2b) impliziert (2.3.2c).
Wird dies unter dem Gesichtspunkt der Frequenz betrachtet (siehe Kapitel 1.1), ist im hochfrequenten Fall die parabolische Zulässigkeitsbedingung die restriktivere, im niedrigfrequenten Fall greift die Standardzulässigkeitsbedingung. Während die Standardzulässigkeitsbedingung meist ein verhältnismäßig kleines Gebiet in der Umgebung von Q_t zu den unzulässigen Gebieten zählt, dargestellt als weißer Bereich in Abbildung 2.5 (a), schrumpft das potentiell zulässige Gebiet enorm durch die Hinzunahme der Richtungen (siehe Abb. 2.5 (b)). Ebenso kann die parabolische Zulässigkeitsbedingung das Gebiet deutlich verkleinern. Die hier Anwendung findenden Zulässigkeitsbedingungen sind damit potentiell erheblich restriktiver als im Standardfall der $\mathcal{H}^2$-Matrizen.
In der Praxis werden die Parameter η_1, η_2 dazu genutzt, die Genauigkeit sowie den Speicherbedarf einzustellen und auf diese Weise eine gute Balance zwischen den beiden Größen zu erhalten.

Aus der Sicht des Praktikers stellt die Wahl der Menge der Richtungen eine kleine Zwickmühle da. Während für eine hohe Güte der Approximation viele Richtungen von Nöten sind, wäre für den Aufwand von Algorithmen eine geringe Anzahl an Richtungen erstrebenswert. Daher sollten so viele Richtungen wie nötig, aber so wenige wie möglich konstruiert wer-

den. Insbesondere der günstigste Fall, dass nur die Null-Richtung $c = 0$ ausreichend ist, sollte schnell erkannt werden. Daher ist es sinnvoll,

$$\kappa \max \{\operatorname{diam}(Q_t), \operatorname{diam}(Q_s)\} \leq \eta_1 \tag{2.3.3}$$

in der Praxis als Erstes zu überprüfen, und dann $c = 0$ zu setzen [3].
Der sichtlich andere Charakter der ersten Zulässigkeitsbedingung (2.3.2a) kann aber auch genutzt werden, um sie im Vorwege sicherzustellen. Nach der Wahl von η_1 kann für jede auftretende Kombination von Gebieten Q_t, Q_s eine hinreichend umfassende Menge an Richtungen $\mathcal{R}$ konstruiert werden. Demzufolge stellt die erste Bedingung nicht direkt eine Einschränkung für die Zulässigkeit dar und kann in der Praxis separat behandelt werden.

Nicht alle betrachteten Kombinationen von Gebieten werden die Zulässigkeitsbedingungen erfüllen können, so dass es notwendig wird, die Approximation auf Teilmatrizen einzuschränken. Um die Konstruktion von Teilmatrizen und den dazugehörigen Gebieten geht es den folgenden zwei Kapiteln.

2.4 Clusterbäume

Um eine Aufteilung einer Matrix $G \in \mathbb{K}^{n \times m}$ in Teilmatrizen zu erhalten, werden Baumstrukturen, die mit einer hierarchischen Aufteilung der Indexmenge assoziiert sind, verwendet. Die Baumstruktur für die Gesamtmatrix setzt sich aus zwei sogenannten *Clusterbäumen* für die Spalten und Zeilen zusammen. Entsprechend soll zunächst das Konzept des Clusterbaums (vgl. [4, K. 3.1]) erläutert werden.

Bei einem Clusterbaum handelt es sich um eine Datenstruktur, die aus hierarchisch angeordneten Knoten, sogenannten *Clustern*, besteht, welche jeweils Verweise zu untergeordneten Clustern enthalten können. Wenn untergeordnete Cluster vorhanden sind, werden diese als *Kinder* bezeichnet, die verweisenden Knoten als *Elterncluster*.
Der Clusterbaum $\mathcal{T}_\mathcal{I}$ wird dabei zu einer endlichen (aber nicht notwendigerweise geordneten) Indexmenge $\mathcal{I} \subset \mathbb{N}$ erstellt. Im Kontext der Matrix stellt die Indexmenge die Menge ihrer Spalten- oder ihrer Zeilenindizes dar. Der Ursprungscluster wird auch *Wurzel* genannt und durch $\operatorname{wurzel}(\mathcal{T}_\mathcal{I})$ gekennzeichnet, er ist mit der vollständigen Indexmenge assoziiert. Der Wurzelcluster wird nach einer festen Vorschrift in disjunkte Kindercluster aufgeteilt, zum Beispiel beim Halbieren in t_1, t_2, und alle Kindercluster eines Clusters t in der Menge $\operatorname{kind}(t)$ zusammengefasst. Entsprechend wird auch mit der assoziierten Indexmenge verfahren, ohne dabei Indizes zu verlieren. Um Verwechselungen vorzubeugen, werden die Indizes des Clusters t mit ${}^\mathfrak{I}t$ bezeichnet. Cluster, deren Indexmenge leer ist, also für die ${}^\mathfrak{I}t = \emptyset$ gilt, sind nicht zugelassen. Bei Bedarf werden Kindercluster erneut unterteilt. Dieses rekursive Vorgehen wird so lange wiederholt, bis die Cluster und damit die entsprechenden Indexmengen klein genug sind. Nicht weiter unterteilte Cluster t werden als *Blätter* bezeichnet,

es gilt $\mathrm{kind}(t) = \emptyset$. Das so entstehende Objekt ist ein *Clusterbaum* zur Indexmenge $\mathcal{I}$ [4, Def. 3.4].

Definition 2.13 (Clusterbaum)
Sei eine endliche Indexmenge $\mathcal{I}$ gegeben. Bezeichne einen Baum $\mathcal{T}_\mathcal{I}$ als Clusterbaum *zur Indexmenge $\mathcal{I}$, falls die Indexmenge zu jedem Cluster $t \in \mathcal{T}_\mathcal{I}$ ein Element der Potenzmenge $\mathcal{P}(\mathcal{I})$ ist und Folgendes erfüllt*

i. für $t = \mathrm{wurzel}(\mathcal{T}_\mathcal{I})$ gilt $^\mathfrak{I}t = \mathcal{I}$,

ii. für alle $t \in \mathcal{T}_\mathcal{I}$ mit $\mathrm{kind}(t) \neq \emptyset$ gilt $^\mathfrak{I}t = \bigcup_{t' \in \mathrm{kind}(t)} {}^\mathfrak{I}t'$ und

iii. für $t \in \mathcal{T}_\mathcal{I}$ mit $t_1, t_2 \in \mathrm{kind}(t)$ gilt entweder $t_1 = t_2$ oder $^\mathfrak{I}t_1 \cap {}^\mathfrak{I}t_2 = \emptyset$.

Bemerkung 8 (Echte Unterteilung): *Die Definition des Clusterbaums lässt zu, dass die Indexmenge eines Clusters t mit der seines Kinds $t' \in \mathrm{kind}\,(t)$ übereinstimmt, also $^\mathfrak{I}t = {}^\mathfrak{I}t'$ gilt. Dies hat jedoch für die hier betrachtete Anwendung keinen Nutzen, im Gegenteil erschwert es, theoretische Aussagen zu treffen. Daher sei im Folgenden immer vorausgesetzt, dass für einen gegebenen Clusterbaum $\mathcal{T}_\mathcal{I}$*

$$\#\,\mathrm{kind}(t) \neq 1 \qquad \textit{für alle } t \in \mathcal{T}_\mathcal{I}$$

gilt.

Je nachdem, wie oft Cluster unterteilt werden, entstehen unterschiedlich viele Hierarchiestufen. Entsprechend lassen sich zu einem Clusterbaum die Begriffe der *Stufe* und *Baumtiefe* definieren [4, Def. 3.6, 3.11].

Definition 2.14 (Stufe & Baumtiefe)
Seien eine endliche Indexmenge $\mathcal{I}$ sowie ein dazugehöriger Clusterbaum $\mathcal{T}_\mathcal{I}$ gegeben. Für jeden Cluster $t \in \mathcal{T}_\mathcal{I}$ sei seine Stufe *gegeben durch*

$$\mathrm{stufe}(t) = \begin{cases} 0 & \textit{falls } t = \mathrm{wurzel}(\mathcal{T}_\mathcal{I}), \\ \mathrm{stufe}(t^+) + 1 & \textit{falls ein } t^+ \in \mathcal{T}_\mathcal{I} \textit{ existiert mit } t \in \mathrm{kind}(t^+). \end{cases}$$

Die maximale Stufe des Clusterbaums $\mathcal{T}_\mathcal{I}$ bezeichne als Baumtiefe

$$p_\mathcal{I} := \max\{\mathrm{stufe}(t) \mid t \in \mathcal{T}_\mathcal{I}\}.$$

Die Cluster eines Clusterbaums $\mathcal{T}_\mathcal{I}$ mit Stufe $\ell \in \underline{p_\mathcal{I}}\,_0$ sind in der Menge $\mathcal{T}_\mathcal{I}^\ell$ vereint. In Abbildung 2.6 ist ein Clusterbaum mit der Indexmenge $\{3, 1, 7, 4, 6\}$ als Wurzel und mit der Baumtiefe $p_\mathcal{I} = 2$ zu sehen. Zudem sind die einzelnen Cluster einer Stufe $\mathcal{T}_\mathcal{I}^\ell$ durch Boxen markiert. Beim Erstellen des Clusterbaums wurde ein Cluster, das aus zwei oder weniger

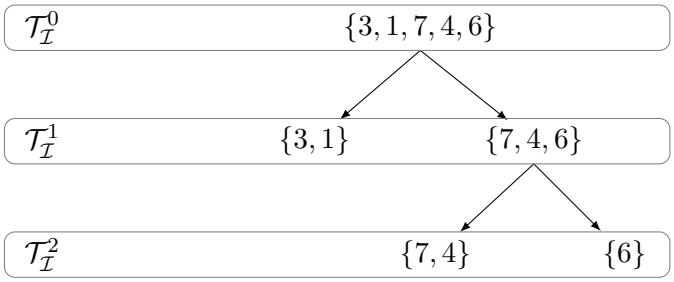


Abbildung 2.6: Simpler Clusterbaum

Elementen besteht, nicht weiter unterteilt. In der Abbildung ist die Hierarchie gut zu erkennen, die sich in verschiedenen Stufen des Clusterbaums und Partitionen von Indexmengen widerspiegelt.

Korollar 2.15 (Endliche Baumtiefe)
Sei ein Clusterbaum $\mathcal{T}_{\mathcal{I}}$ zur Indexmenge $\mathcal{I}$ mit $n = \#\mathcal{I}$ gegeben. Aus der Endlichkeit der Indexmenge folgt mit Bemerkung 8 direkt, dass $p_{\mathcal{I}} \leq n$.

Beweis: Für alle $t \in \mathcal{T}_{\mathcal{I}}$ und $t' \in \operatorname{kind}(t)$ gilt unabhängig von der verwendeten Aufteilungsstrategie wegen $\#\operatorname{kind}(t) \neq 1$ immer $\#^{\mathfrak{I}}t' < \#^{\mathfrak{I}}t$. Da zusätzlich keine Cluster mit leerer Indexmenge zugelassen sind, folgt aus $n = \#\mathcal{I}$ direkt, dass maximal n Stufen vorhanden sein können. □

Eine wichtige Rolle spielt die Menge der Cluster, die keine Kinder mehr haben, diese Cluster werden in der Menge der *Blätter* zusammengefasst [4, S. 30].

Definition 2.16 (Blätter)
Sei ein Clusterbaum $\mathcal{T}_{\mathcal{I}}$ zur Indexmenge $\mathcal{I}$ gegeben, bezeichne die durch

$$\mathcal{L}_{\mathcal{I}} := \{t \in \mathcal{T}_{\mathcal{I}} \mid \operatorname{kind}(t) = \emptyset\}$$

gegebene Menge als Blätter *des Clusterbaums $\mathcal{T}_{\mathcal{I}}$.*

Es kommt vor, dass eine Operation nicht nur auf die Kinder, sondern auch auf die Kinder der Kinder und deren Kinder angewendet werden muss. Dies führt zum Konzept der *Nachfahren* eines Clusters $t \in \mathcal{T}_{\mathcal{I}}$. Soll etwas in der Hierarchie weiter nach oben getragen werden, ist entsprechend eine Menge der *Vorfahren* von Nöten. Daher sollen auch für diese Mengen Bezeichnungen eingeführt werden [4, Def. 3.5].

Definition 2.17 (Nach- und Vorfahren)
Sei ein Clusterbaum $\mathcal{T}_{\mathcal{I}}$ zur Indexmenge $\mathcal{I}$ gegeben. Für einen Cluster $t \in \mathcal{T}_{\mathcal{I}}$ ist die Menge

seiner Nachfahren *gegeben durch:*

$$\mathrm{nac}(t) := \begin{cases} \{t\} & \textit{falls } \mathrm{kind}(t) = \emptyset, \\ \{t\} \cup \bigcup\limits_{t' \in \mathrm{kind}(t)} \mathrm{nac}(t') & \textit{sonst.} \end{cases}$$

Die Menge der Vorfahren *des Clusters $t \in \mathcal{T}_\mathcal{I}$ ist durch*

$$\mathrm{vor}(t) := \left\{ t^+ \in \mathcal{T}_\mathcal{I} \mid t \in \mathrm{nac}(t^+) \right\}$$

gegeben.

Bemerkung 9 : *Um Algorithmen leichter und kürzer hinschreiben zu können, ist der betrachtete Cluster selbst ein Teil seiner Vor- und Nachfahren.*

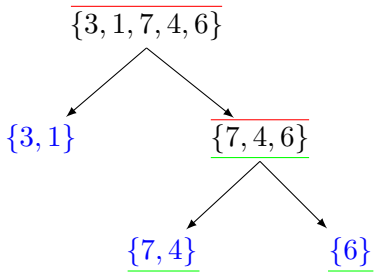


Abbildung 2.7: Clusterbaum mit markierten Vorfahren, Nachfahren und Blättern

In der Abbildung 2.7 sind die Blattcluster in blau geschrieben, die Vorfahren des Clusters $\{7, 4, 6\}$ sind mit einem roten Strich und die Nachfahren desselben Clusters mit einem grünen Strich gekennzeichnet. Ein weiteres komfortables Hilfsmittel stellt das Konzept der Teilbäume dar (vgl. [32, Def. A.3.1]).

Definition 2.18 (Teilbaum)
Seien ein Clusterbaum $\mathcal{T}_\mathcal{I}$ und $t \in \mathcal{T}_\mathcal{I}$ gegeben. Definiere den zu t gehörenden Teilbaum $\mathcal{T}_{\mathfrak{I}_t}$ durch den mit Wurzel t gebildeten Teil des Clusterbaums $\mathcal{T}_\mathcal{I}$. Das heißt, dass für $t' \in \mathcal{T}_\mathcal{I}$ mit $t' \in \mathrm{nac}(t)$ auch $t' \in \mathcal{T}_{\mathfrak{I}_t}$ gilt, ebenso bleibt die Hierarchie von $\mathcal{T}_\mathcal{I}$ erhalten, also falls $s \in \mathrm{kind}(t')$ in $\mathcal{T}_\mathcal{I}$ erfüllt, so gilt dies auch in $\mathcal{T}_{\mathfrak{I}_t}$. Synonym nutze auch die Schreibweise $\mathcal{T}_t$ für den Teilbaum $\mathcal{T}_{\mathfrak{I}_t}$.

Jedoch reichen Clusterbäume allein für die Helmholtz-Gleichung nicht aus. Für die Modifizierung der Kernfunktionen werden noch zu jedem Cluster Richtungen benötigt. Entsprechend definiere zunächst die Menge der Richtungen eines Clusters [3, Def. A.4].

Definition 2.19 (Menge der Richtungen)
Definiere zu einem gegebenen Clusterbaum $\mathcal{T}_\mathcal{I}$ die Menge der zugehörigen Richtungen

$\mathcal{R} := \{\mathcal{R}_\ell\}_{\ell \in \underline{p_\mathcal{I}}_{\rfloor 0}}$ als Familie von Teilmengen des $\mathbb{R}^d$, wobei eine einzelne Richtung $c \in \mathcal{R}_\ell$ für $\ell \in \underline{p_\mathcal{I}}_{\rfloor 0}$ entweder

$$\|c\|_2 = 1 \qquad \textit{oder} \qquad c = 0$$

erfüllt. Zu jeder Stufe $\ell \in \underline{p_\mathcal{I} - 1}_{\rfloor 0}$ existiert eine Abbildung $\mathrm{r}_\ell : \mathcal{R}_\ell \to \mathcal{R}_{\ell+1}$, die jeder Richtung $c \in \mathcal{R}_\ell$ ihre Bestapproximation in der Menge $\mathcal{R}_{\ell+1}$ zuordnet, entsprechend gilt

$$\|c - \mathrm{r}_\ell(c)\|_2 \leq \|c - c'\|_2 \qquad \textit{für alle } c' \in \mathcal{R}_{\ell+1}.$$

Für $\ell \in \underline{p_\mathcal{I} - 1}_{\rfloor 0}$ und eine Richtung $c \in \mathcal{R}_{\ell+1}$ sei durch $\mathrm{r}_\ell^{-1}(c)$ das Urbild von c unter r_ℓ gegeben.
Synonym verwende die Bezeichnung $\mathcal{R}_t$ für die Richtungen von einem Cluster $t \in \mathcal{T}_\mathcal{I}$ als kürzere Schreibweise von $\mathcal{R}_{\mathrm{stufe}(t)}$ und ebenso r_t anstelle von r_ℓ.

Es sei bemerkt, dass zusätzlich die Richtung $c = 0$ hinzugefügt wurde, um den niederfrequenten Fall ohne die hier unnötige Modifizierung der Kernfunktion betrachten zu können.

Durch die Kombination eines Clusterbaums und einer dazugehörigen Menge von Richtungen wird ein *richtungsabhängiger Clusterbaum* konstruiert.

Definition 2.20 (Richtungsabhängiger Clusterbaum)
Ein richtungsabhängiger Clusterbaum *zu einer Indexmenge $\mathcal{I}$ besteht aus einem Clusterbaum $\mathcal{T}_\mathcal{I}$ und einer dazugehörigen Familie von Richtungsmengen $\mathcal{R}$. Entsprechend steht jedem Cluster $t \in \mathcal{T}_\mathcal{I}$ eine Auswahl an möglichen Richtungen $c \in \mathcal{R}_t$ zur Verfügung.*

Die bisher eingeführten Begriffe und vorgestellten Eigenschaften von Clusterbäumen sind richtungsunabhängig und übertragen sich daher direkt auf die richtungsabhängige Variante. Unter Zuhilfenahme des folgenden Lemmas lässt sich die praktische Partitionseigenschaft von Clusterbäumen zeigen.

Lemma 2.21 (Schnitte von Clustern)
Seien ein Clusterbaum $\mathcal{T}_\mathcal{I}$ und zwei Cluster $t, s \in \mathcal{T}_\mathcal{I}$ mit ${}^\mathfrak{I}t \cap {}^\mathfrak{I}s \neq \emptyset$ und $\mathrm{stufe}(t) \leq \mathrm{stufe}(s)$ gegeben, dann gilt entweder

$$\begin{aligned} &t = s && \textit{falls } \mathrm{stufe}(t) = \mathrm{stufe}(s) && \textit{oder} \\ &s \in \mathrm{nac}(t) && \textit{falls } \mathrm{stufe}(t) < \mathrm{stufe}(s). \end{aligned}$$

Ein Beweis des Lemmas findet sich in [4, S. 33].

Lemma 2.22 (Stufen- & Blattpartition)
Die Menge der Blätter eines Clusterbaums $\mathcal{T}_\mathcal{I}$ bildet eine Partition der Indexmenge $\mathcal{I}$.
Die Menge der Cluster $\mathcal{T}_\mathcal{I}^\ell$ mit Stufe $\ell \in \underline{p_\mathcal{I}}_{\rfloor 0}$ bildet eine Partition einer Teilmenge der Indexmenge $\mathcal{I}$.

Ein Beweis des Lemmas findet sich in [4, S. 33 f.].

Mit der Blattpartition ist es möglich, Abschätzungen für $\mathcal{T}_\mathcal{I}$ und $\mathcal{L}_\mathcal{I}$ anzugeben, was wiederum nötig ist, um konkrete Komplexitäten zu bestimmen.

Lemma 2.23 (Cluster)
Sei zur Indexmenge $\mathcal{I}$ ein Clusterbaum $\mathcal{T}_\mathcal{I}$ gegeben. Die Menge aller im Baum auftretenden Cluster ist beschränkt mit

$$\#\mathcal{T}_\mathcal{I} \leq 2\#\mathcal{L}_\mathcal{I} - 1.$$

Ein Beweis für das Lemma befindet sich in [32, S. 358-359].

Ohne weitere Annahmen kann eine triviale Abschätzung für die Anzahl der Blätter aus dem Lemma 2.22 zur Blattpartition gefolgert werden.

Korollar 2.24 (Anzahl Blätter)
Sei zur Indexmenge $\mathcal{I}$ ein Clusterbaum $\mathcal{T}_\mathcal{I}$ gegeben. Die Anzahl der Blätter des Clusterbaums $\mathcal{T}_\mathcal{I}$ ist dann beschränkt durch

$$\#\mathcal{L}_\mathcal{I} \leq n.$$

Für die bisher betrachtete Theorie reichte eine bloße Indexmenge aus. In der Praxis steht diese Indexmenge nicht für sich allein, vielmehr existieren Geometrien, die mit den Indizes korrespondieren, so zum Beispiel Dreiecke aus Oberflächentriangulationen oder Punktmengen aus Gittern. Es existiert also ein Gebiet Ω, das mit der Wurzel $\mathrm{wurzel}(\mathcal{T}_\mathcal{I})$ korrespondiert. Entsprechend gibt es Teilgebiete Ω_t, die zu den einzelnen Clustern t in Relation stehen.

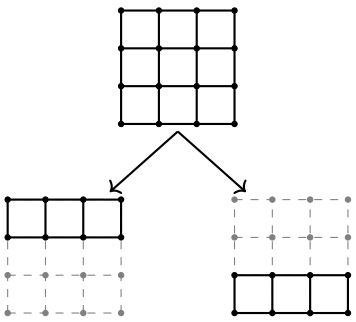


Abbildung 2.8: Geometrisches Äquivalent zum Clusterbaum der 0.ten und 1.ten Stufe

Die Abbildung 2.8 zeigt ein quadratisches Gitter mit 16 Punkten als Wurzelcluster, welches zwei Kindercluster mit jeweils 8 Punkten hat.

Im Hinblick auf die spätere Approximation ist es zudem wichtig, dass der Träger der Basisfunktion zu einem Index vollständig im dazugehörigen Gebiet enthalten ist und nicht nur das geometrische Gegenstück zum Index selbst. Je nach Wahl der Basisfunktionen kommt es so zu Überlappungen benachbarter Gebiete.
Treten zum Beispiel Punkte gehäuft auf oder ist die Grundgeometrie komplexer, entstehen schnell Gebiete, deren Handhabung deutlich schwieriger ist. Schon die algorithmische Bestimmung der Größe eines allgemeinen Polyeders im Dreidimensionalen ist aufwendig. Aus diesem Grund wird zu jedem Cluster t ein achsenparalleler *überdeckender Quader* Q_t eingeführt. Quader sind leicht zu konstruieren und Abstände sowie Durchmesser gut zu berechnen (siehe Kapitel 2.3).

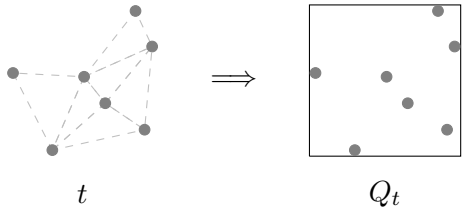


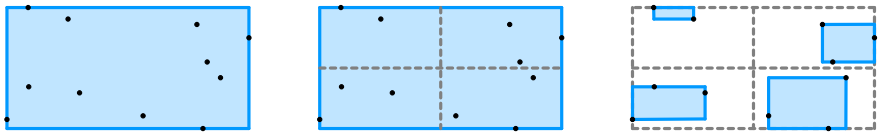
Abbildung 2.9: Überdeckender Quader zu einem Cluster t

Die Abbildung 2.9 zeigt den kleinsten überdeckenden Quader zum Cluster t, der sechs Dreiecke einer unregelmäßigen Triangulation enthält. Dabei wurde bei der Bildung des überdeckenden Quaders angenommen, dass stückweise konstante Basisfunktion auf den Dreiecken verwendet werden, so dass es ausreicht, die Dreiecke vollständig zu überdecken.

Geometrische Informationen erlauben es, einen genaueren Blick auf Clusterstrategien zu werfen, was wiederum schärfere Abschätzungen für die Baumtiefe erlaubt. Zwei typische Strategien sollen hier vorgestellt werden. Da Halbieren beim Aufteilen meist ausreicht, so auch bei der Helmholtz-Gleichung, wird dies im Folgenden betrachtet.

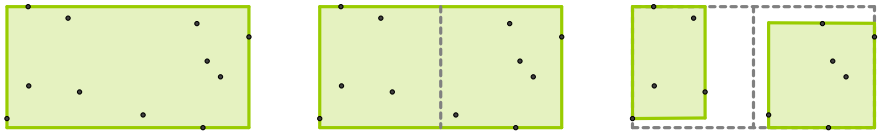
Bemerkung 10 (Simultane Clusterstrategie): *Bei der simultanen Clusterstrategie wird in jedem Schritt entlang jeder Ausdehnungsdimension unterteilt. Im d-Dimensionalen hat ein unterteilter Cluster beim Halbieren somit maximal 2^d Kinder. Potentielle Kinder, die keine Elemente enthalten, werden weggelassen. Dies hat zur Folge, dass die Anzahl der Cluster mit wachsender Stufe schnell steigt, jedoch schrumpfen die Durchmesser der überdeckenden Quader auch entsprechend rapide.*

Die Abbildung 2.10 zeigt, ausgehend von einem gegebenen Quader (a), die Aufteilung in Kindercluster (b) und die zu den Kinderclustern gehörenden minimalen überdeckenden Quader (c) bei der simultanen Clusterstrategie. Die hier gebildeten überdeckenden Quader beziehen sich ebenfalls allein auf die Punkte im Cluster.



(a) Ausgangsquader (b) Aufteilung (c) Minimale überdeckende Quader

Abbildung 2.10: Simultane Unterteilung



(a) Ausgangsquader (b) Aufteilung (c) Minimale überdeckende Quader

Abbildung 2.11: Adaptive Unterteilung

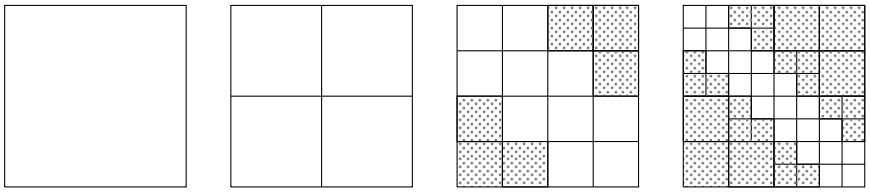
Bemerkung 11 (Adaptive Clusterstrategie): *In der Praxis hat sich die adaptive Clusterstrategie bewährt, bei der in jedem Schritt entlang der größten Ausdehnungsdimension unterteilt wird. Zwar schrumpfen die Durchmesser nicht mehr so schnell, was zu höheren Baumtiefen führen kann, dafür aber kommen beim Halbieren nur* 2 *Kinder auf einen unterteilten Cluster.*

Abbildung 2.11 zeigt das Vorgehen bei der adaptiven Clusterstrategie an demselben Cluster wie in Abbildung 2.10 zur simultanen Clusterstrategie. Auch hier wurden die überdeckenden Quader allein um die Punkte im Cluster gebildet, ohne möglicherweise überstehende Träger von Basisfunktionen zu beachten.

Bemerkung 12 (Baumtiefe): *Mit zusätzlichen Voraussetzungen an das Gitter und Grenzen für die minimale Größe von Clustern und damit einer unteren Grenze für die minimale Ausdehnung der überdeckenden Quader, kann für die vorgestellten Clusterstrategien* $p_{\mathcal{I}} \sim \log_2(\#\mathcal{I})$ *gezeigt werden [5, S. 39 ff.] [6, Bem. 13] .*

2.5 Blockbäume

Zum Beschreiben einer Matrix werden Zeilen und Spalten benötigt, daher ist es notwendig, den Clusterbaum zum *Blockbaum* zu erweitern. Für den Blockbaum $\mathcal{T}_{\mathcal{I}\times\mathcal{I}}$ wird eine Teilmenge des kartesischen Produkts des Clusterbaums $\mathcal{T}_{\mathcal{I}}$ mit sich selbst verwendet. In Bezug auf die Matrix bedeutet dies die Nutzung eines Zeilen- und eines Spaltenclusterbaums, um



(a) Stufe 0 (b) Stufe 1 (c) Stufe 2 (d) Stufe 3

Abbildung 2.12: Erste Stufen eines möglichen Blockbaums

die Matrix in Teilmatrizen zu zerlegen.
Für die Approximation sind zusätzlich noch Richtungen für die ebenen Wellen notwendig, so dass einem Element des Blockbaums eine eindeutige Richtung zugeordnet wird. Entsprechend muss der verwendete Clusterbaum richtungsabhängig sein.

Definition 2.25 (Richtungsabhängiger Blockbaum)
Seien ein richtungsabhängiger Clusterbaum $\mathcal{T}_\mathcal{I}$ zur Indexmenge $\mathcal{I}$ und eine dazugehörige Familie an Richtungsmengen $\mathcal{R}$ gegeben. Bezeichne $\mathcal{T}_{\mathcal{I}\times\mathcal{I}}$ als richtungsabhängigen Blockbaum zur Indexmenge $\mathcal{I}\times\mathcal{I}$, falls Folgendes gilt:

i. Für jeden Block $b \in \mathcal{T}_{\mathcal{I}\times\mathcal{I}}$ existieren Cluster $t, s \in \mathcal{T}_\mathcal{I}$ mit $\mathrm{stufe}(t) = \mathrm{stufe}(s)$ und eine Richtung $c_b \in \mathcal{R}_t$, die zusammen $b = (t, s, c_b)$ erfüllen.

ii. Die Wurzel $r = \mathrm{wurzel}(\mathcal{T}_{\mathcal{I}\times\mathcal{I}})$ erfüllt $r = (\mathrm{wurzel}(\mathcal{T}_\mathcal{I}), \mathrm{wurzel}(\mathcal{T}_\mathcal{I}), c_r)$ mit einem $c_r \in \mathcal{R}_0$.

iii. Für jedes $b \in \mathcal{T}_{\mathcal{I}\times\mathcal{I}}$ mit $b = (t, s, c_b)$ ist die dazugehörige Indexmenge mit ${}^\mathfrak{I}b = {}^\mathfrak{I}t \times {}^\mathfrak{I}s$ gegeben.

iv. Falls es sich bei $b = (t, s, c_b) \in \mathcal{T}_{\mathcal{I}\times\mathcal{I}}$ um kein Blatt handelt, gilt $\mathrm{kind}(t), \mathrm{kind}(s) \neq \emptyset$. Zu jedem $b' \in \mathrm{kind}(b)$ existieren $t' \in \mathrm{kind}(t), s' \in \mathrm{kind}(s)$ sowie $c'_b \in \mathcal{R}_{t'}$ mit $b' = (t', s', c'_b)$ und es gilt ${}^\mathfrak{I}b = \dot{\bigcup}_{b'\in\mathrm{kind}(b)} {}^\mathfrak{I}b'$.

Über die erste Bedingung in der Definition des Blockbaums ist sichergestellt, dass jedem Block auch eine Richtung c_b zugeordnet werden kann. Denn erst die Einschränkung, dass die Cluster t und s eines Blocks $\mathrm{stufe}(t) = \mathrm{stufe}(s)$ erfüllen müssen, sorgt dafür, dass $\mathcal{R}_t = \mathcal{R}_s$ gilt.

Die Abbildung 2.12 zeigt eine Zerlegung einer Matrix mit Hilfe eines Blockbaums. Ganz links in (a) ist die Ursprungsmatrix zu sehen, (b) zeigt die Blöcke auf der ersten Stufe. Auch in diesem Fall wurden die Cluster halbiert, so dass vier Blöcke als Kinder erhalten werden. Für (c) und (d) wurden die Cluster und damit die Blöcke weiter unterteilt. Grau gekennzeichnete Blöcke werden nicht mehr weiter unterteilt.

Wie schon bei den Clusterbäumen sollen auch für Blockbäume einige wichtige Begriffe und Mengen eingeführt werden. Die Konzepte der *Stufe*, *Baumtiefe* und *Blätter* lassen sich leicht auf den richtungsabhängigen Blockbaum übertragen.

Definition 2.26 (Stufe, Baumtiefe und Blätter)
Sei ein richtungsabhängiger Blockbaum $\mathcal{T}_{\mathcal{I}\times\mathcal{I}}$ gegeben. Die Stufe eines Blocks ist durch die Stufe der zugehörigen Cluster definiert

$$\operatorname{stufe}(b) := \operatorname{stufe}(t) = \operatorname{stufe}(s) \qquad \textit{für alle } b = (t, s, c_b) \in \mathcal{T}_{\mathcal{I}\times\mathcal{I}},$$

die Baumtiefe $p_{\mathcal{I}\times\mathcal{I}}$ des Blockbaums ist das Maximum der auftretenden Stufen

$$p_{\mathcal{I}\times\mathcal{I}} := \max\{\operatorname{stufe}(b) \mid b \in \mathcal{T}_{\mathcal{I}\times\mathcal{I}}\} \leq p_{\mathcal{I}}.$$

Blöcke ohne Kinder werden als Blätter bezeichnet und in der Menge

$$\mathcal{L}_{\mathcal{I}\times\mathcal{I}} := \{b \in \mathcal{T}_{\mathcal{I}\times\mathcal{I}} \mid \operatorname{kind}(b) = \emptyset\}$$

vereinigt.

Ebenso lassen sich die Konzepte der *Vor-* und *Nachfahren* übertragen.

Definition 2.27 (Vor- und Nachfahren)
Die Nachfahren *eines Blocks $b \in \mathcal{T}_{\mathcal{I}\times\mathcal{I}}$ sind definiert durch*

$$\operatorname{nac}(b) := \begin{cases} \{b\} & \textit{falls } \operatorname{kind}(b) = \emptyset, \\ \{b\} \cup \bigcup\limits_{b' \in \operatorname{kind}(b)} \operatorname{nac}(b') & \textit{sonst} \end{cases}$$

und die Vorfahren *eines Blocks $b \in \mathcal{T}_{\mathcal{I}\times\mathcal{I}}$ durch*

$$\operatorname{vor}(b) := \left\{b^+ \in \mathcal{T}_{\mathcal{I}\times\mathcal{I}} \mid b \in \operatorname{nac}(b^+)\right\}.$$

Wird die Richtung eines Blocks als sekundäre Eigenschaft betrachtet, kann der richtungsabhängige Blockbaum als Clusterbaum zur Indexmenge $\mathcal{I} \times \mathcal{I}$ wie in [4, S. 35] interpretiert werden.

Lemma 2.28 (Clusterbaum zur Indexmenge $\mathcal{I} \times \mathcal{I}$)
Ein richtungsabhängiger Blockbaum $\mathcal{T}_{\mathcal{I}\times\mathcal{I}}$ ist ein Clusterbaum zur Indexmenge $\mathcal{I} \times \mathcal{I}$.

Beweis: Führe den Beweis durch Überprüfen der Definition eines Clusterbaums.
Sei $b_r \in \mathcal{T}_{\mathcal{I}\times\mathcal{I}}$ die Wurzel des richtungsabhängigen Blockbaums. Dann gilt nach der Definition $b_r = (\operatorname{wurzel}(\mathcal{T}_{\mathcal{I}}), \operatorname{wurzel}(\mathcal{T}_{\mathcal{I}}), c_r)$ und damit folgt direkt aus $\operatorname{wurzel}(\mathcal{T}_{\mathcal{I}}) = \mathcal{I}$,

dass ${}^{\mathfrak{I}}b_r = \mathcal{I} \times \mathcal{I}$ erfüllt. Sei ein $b = (t, s, c_b) \in \mathcal{T}_{\mathcal{I}\times\mathcal{I}}$ mit $\mathrm{kind}(b) \neq \emptyset$ gegeben. Es gilt nach der Definition des Blockbaums

$$ {}^{\mathfrak{I}}b = \bigcup_{b' \in \mathrm{kind}(b)} {}^{\mathfrak{I}}b'. $$

Für die letzte Eigenschaft seien $b_1, b_2 \in \mathrm{kind}(b)$ mit $b_1 = (t_1, s_1, c_1)$, $b_2 = (t_2, s_2, c_2)$ und $b_1 \neq b_2$ gegeben. Dann gilt

$$ {}^{\mathfrak{I}}b_1 \cap {}^{\mathfrak{I}}b_2 = ({}^{\mathfrak{I}}t_1 \times {}^{\mathfrak{I}}s_1) \cap ({}^{\mathfrak{I}}t_2 \times {}^{\mathfrak{I}}s_2) = ({}^{\mathfrak{I}}t_1 \cap {}^{\mathfrak{I}}t_2) \times ({}^{\mathfrak{I}}s_1 \cap {}^{\mathfrak{I}}s_2). $$

Da $b_1 \neq b_2$ und $t_1, t_2 \in \mathrm{kind}(t)$, $s_1, s_2 \in \mathrm{kind}(s)$ erfüllen, gilt nach Definition des Clusterbaums $t_1 \neq t_2$ oder $s_1 \neq s_2$ und damit ${}^{\mathfrak{I}}t_1 \cap {}^{\mathfrak{I}}t_2 = \emptyset$ oder ${}^{\mathfrak{I}}s_1 \cap {}^{\mathfrak{I}}s_2 = \emptyset$. Somit folgt dann ${}^{\mathfrak{I}}b_1 \cap {}^{\mathfrak{I}}b_2 = \emptyset$. □

Dies ermöglicht es, weitere Eigenschaften von Clusterbäumen zu übertragen. Ein für die Approximation besonders wichtiges Resultat ist die Blattpartition.

Lemma 2.29 (Blattpartition)
Sei ein richtungsabhängiger Blockbaum $\mathcal{T}_{\mathcal{I}\times\mathcal{I}}$ gegeben. Die Menge seiner Blätter $\mathcal{L}_{\mathcal{I}\times\mathcal{I}}$ bildet eine Partition der Indexmenge $\mathcal{I} \times \mathcal{I}$.

Beweis: Nach Lemma 2.28 handelt es sich bei dem richtungsabhängigen Blockbaum um einen Clusterbaum und das Lemma 2.22 liefert die Blattpartition für Clusterbäume. □

Beim Beschreiben von Algorithmen mit Blockbäumen wird öfter die Menge der Spaltencluster, die mit einem speziellen Zeilenclusters $t \in \mathcal{T}_{\mathcal{I}}$ zusammen auftauchen, beziehungsweise die Menge der Zeilencluster, die zusammen mit einem speziellen Spaltencluster $s \in \mathcal{T}_{\mathcal{I}}$ auftauchen, verwendet (vgl. [4, Def. 3.29]).

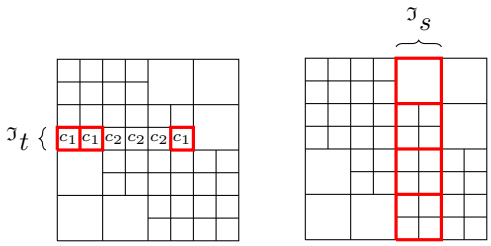
Definition 2.30 (Clusterpartner)
Sei ein richtungsabhängiger Blockbaum $\mathcal{T}_{\mathcal{I}\times\mathcal{I}}$ zur Indexmenge $\mathcal{I}$ gegeben. Bezeichne zu einem gegebenen Cluster $t \in \mathcal{T}_{\mathcal{I}}$ alle Cluster $s \in \mathcal{T}_{\mathcal{I}}$, für die ein Block $(t, s, c) \in \mathcal{T}_{\mathcal{I}\times\mathcal{I}}$ existiert, als Spaltenclusterpartner *von t. Fasse alle Spaltenclusterpartner, die zusammen mit einer Richtung auftreten, in der Menge*

$$ \mathrm{row}_c(t) := \{s \in \mathcal{T}_{\mathcal{I}} \mid (t, s, c) \in \mathcal{T}_{\mathcal{I}\times\mathcal{I}}\} \qquad \text{für alle } t \in \mathcal{T}_{\mathcal{I}} \text{ und } c \in \mathcal{R}_t $$

zusammen. Analog wird zu einem Cluster $s \in \mathcal{T}_{\mathcal{I}}$ die Menge der Zeilenclusterpartner *definiert, fasse alle Zeilenclusterpartner von s, die zusammen mit einer Richtung auftreten in der Menge*

$$ \mathrm{col}_c(s) := \{t \in \mathcal{T}_{\mathcal{I}} \mid (t, s, c) \in \mathcal{T}_{\mathcal{I}\times\mathcal{I}}\} \qquad \text{für alle } s \in \mathcal{T}_{\mathcal{I}} \text{ und } c \in \mathcal{R}_s $$

zusammen.



(a) $\mathrm{row}_{c_1}(t)$ (b) $\mathrm{col}(s)$

Abbildung 2.13: Beispiele für Zeilen- beziehungsweise Spaltenclusterpartner

Die Vereinigung über alle Richtungen einer Stufe liefert alle Spalten- beziehungsweise Zeilenclusterpartner

$$\mathrm{row}(t) := \bigcup_{c \in \mathcal{R}_t} \mathrm{row}_c(t), \qquad \mathrm{col}(s) := \bigcup_{c \in \mathcal{R}_s} \mathrm{col}_c(s) \qquad \text{für alle } t, s \in \mathcal{T}_{\mathcal{I}}.$$

In der Abbildung 2.13 ist links die Menge $\mathrm{row}_{c_1}(t)$ für das gekennzeichnete Cluster t mit Richtung c_1 rot markiert, während die rechte Abbildung in rot $\mathrm{col}(s)$ zeigt.

Das übergeordnete Ziel lautet, möglichst große Teilmatrizen der Systemmatrix A zu approximieren. Dieses Ziel soll schon beim Bau des Blockbaums berücksichtigt werden, weshalb es ungünstig wäre, Blöcke weiter zu unterteilen, nur weil die zugehörigen Cluster dies zuließen. An dieser Stelle kommen die Zulässigkeitsbedingungen erneut ins Spiel. Da die Zulässigkeit mit der Approximierbarkeit korrespondiert, wurden die für die Helmholtz-Gleichung nötigen Zulässigkeitsbedingungen schon im Kapitel 2.3 hergeleitet.
Da die weitere Unterteilung zulässiger Blöcke aus Gesichtspunkten der Effizienz zu vermeiden ist, ergibt sich der folgende Algorithmus 2.1 zur Erstellung eines richtungsabhängigen Blockbaums. Dabei bezeichne $\mathcal{Z}_s$ die Standardzulässigkeitsbedingung (2.3.2c) und $\mathcal{Z}_p$ die parabolische Zulässigkeitsbedingung (2.3.2b).

Zumeist wird mit den Blättern gearbeitet, da sie eine Partition der Matrix liefern und das Muster für die Speicherung festlegen. Die Zulässigkeitsbedingung wiederum erlaubt eine Partition der Blockblätter in zwei Mengen. Zum einen die zulässigen Blockblätter, die über eine Approximation dargestellt werden sollen, zum anderen die unzulässigen Blockblätter. Die beiden Mengen der Blockblätter werden als *Fern-* und *Nahfeld* bezeichnet.

Definition 2.31 (Nah- und Fernfeld)
Seien ein richtungsabhängiger Blockbaum $\mathcal{T}_{\mathcal{I}\times\mathcal{I}}$ und eine Zulässigkeitsbedingung $\mathcal{Z}$ gegeben. Bezeichne die Menge der zulässigen Blöcke $b \in \mathcal{L}_{\mathcal{I}\times\mathcal{I}}$ als Fernfeld und fasse sie

function build_blocktree($t, s, \mathcal{Z}_s, \mathcal{Z}_p, \mathcal{R}$)
  Boolean z, block b, b', direction c_b
  if $\mathcal{Z}_s(t,s)$ and $\mathcal{Z}_p(t,s)$ are true **then**
    $z \leftarrow$ true
  else
    $z \leftarrow$ false
  end if
  $c_b \leftarrow c \in \mathcal{R}_t$ such that $\left\| \frac{m_t - m_s}{\|m_t - m_s\|_2} - c_b \right\|_2 \leq \left\| \frac{m_t - m_s}{\|m_t - m_s\|_2} - c \right\|_2$ for all $c \in \mathcal{R}_t$
  $b \leftarrow (t, s, c_b)$
  if $z == false$ and $\mathrm{kind}(t) \neq \emptyset$ and $\mathrm{kind}(s) \neq \emptyset$ **then**
    $\mathrm{kind}(b) = \emptyset$
    for all $t' \in \mathrm{kind}(t)$ **do**
      for all $s' \in \mathrm{kind}(s)$ **do**
        $b' \leftarrow$ build_blocktree($t', s', \mathcal{Z}_s, \mathcal{Z}_p, \mathcal{R}$)
        $\mathrm{kind}(b) \leftarrow \mathrm{kind}(b) \cup b'$
      end for
    end for
  end if
  return b
end function

Algorithmus 2.1: Konstruktion des richtungsabhängigen Blockbaums

in

$$\mathcal{L}^+_{\mathcal{I}\times\mathcal{I}} := \{b \in \mathcal{L}_{\mathcal{I}\times\mathcal{I}} \mid \mathcal{Z}(b) = \textit{wahr}\} \subset \mathcal{L}_{\mathcal{I}\times\mathcal{I}}$$

zusammen. Die Menge der unzulässigen Blöcke $b \in \mathcal{L}_{\mathcal{I}\times\mathcal{I}}$ *bilden das* Nahfeld*, welches entsprechend durch* $\mathcal{L}^-_{\mathcal{I}\times\mathcal{I}} := \mathcal{L}_{\mathcal{I}\times\mathcal{I}} \setminus \mathcal{L}^+_{\mathcal{I}\times\mathcal{I}}$ *gegeben ist.*

In Bezug auf die Zulässigkeitsbedingung können die Begriffe der Zeilen- und Spaltenclusterpartner erweitert werden. Dies ist dann von Nutzen, wenn Algorithmen nur mit zulässigen oder unzulässigen Blöcken arbeiten. Weshalb die zulässigen Zeilen- und Spaltenclusterpartner durch

$$\begin{aligned}\mathrm{row}^+(t) &:= \left\{s \in \mathcal{T}_{\mathcal{I}} \mid \exists c \in \mathcal{R}_t \text{ mit } (t,s,c) \in \mathcal{L}^+_{\mathcal{I}\times\mathcal{I}}\right\} \subset \mathrm{row}(t)\\ \mathrm{col}^+(s) &:= \left\{t \in \mathcal{T}_{\mathcal{I}} \mid \exists c \in \mathcal{R}_s \text{ mit } (t,s,c) \in \mathcal{L}^+_{\mathcal{I}\times\mathcal{I}}\right\} \subset \mathrm{col}(s)\end{aligned}$$

gegeben sind.

Für die Theorie reicht es zu wissen, dass jeder Block eine eindeutige Richtung besitzt. Für die Praxis stellt sich die Frage, welche der endlich vielen Richtungen auf der entsprechenden Stufe gewählt werden sollte.
Durch die räumliche Anordnung der beiden Clusterpartner t, s eines richtungsabhängigen Blocks $b \in \mathcal{T}_{\mathcal{I}\times\mathcal{I}}$ zueinander wird die Ausbreitungsrichtung für die ebene Welle definiert. Dabei wird die Welle grundsätzlich als vom Spaltencluster ausgehend angenommen. Da im Allgemeinen jeder Punkt aus t zu jedem Punkt aus s eine andere Richtung für die ebene Welle benötigen könnte, werden die Mittelpunkte der Cluster m_t, m_s zu Hilfe genommen. Mit ihnen kann eine eindeutige Richtung $\hat{c} \in \mathbb{R}^3$ für den Block b festgelegt werden. Jedoch ist durch $\mathcal{R}_t$ die Menge der zur Verfügung stehenden Richtungen für die betrachtete Stufe eingeschränkt. Entsprechend wird die verwendete Richtung $c_b \in \mathcal{R}_t$ so gewählt, dass sie die tatsächliche Richtung $\hat{c}$ am besten approximiert

$$\left\| \tfrac{m_t - m_s}{\|m_t - m_s\|_2} - c_b \right\|_2 \leq \left\| \tfrac{m_t - m_s}{\|m_t - m_s\|_2} - c \right\|_2 \qquad \text{für alle } c \in \mathcal{R}_t.$$

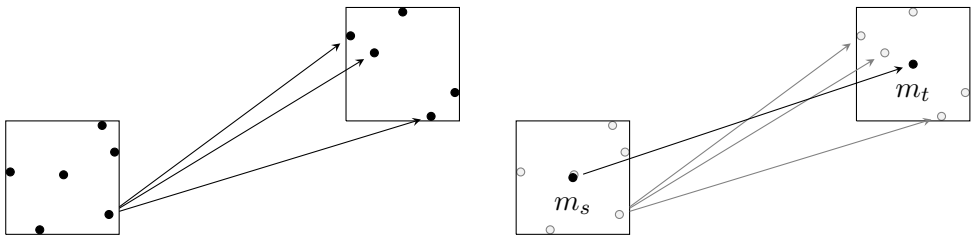


Abbildung 2.14: Festlegung der Richtung eines Blocks $b = (t, s, c_b)$

Anschaulich wird das Vorgehen in Abbildung 2.14 dargestellt. Die Punkte des Clusters t werden durch den Mittelpunkt m_t ersetzt, entsprechend wird auch mit den Richtungen verfahren.
Eine Möglichkeit, um zu gewährleisten, dass für alle Clusterpartner t, s mit $\ell = \text{stufe}(t)$ eine adäquate Richtung $c \in \mathcal{R}_\ell$ vorliegt, ist, die Richtungen so zu konstruieren, dass für alle $z \in \mathbb{R}^3$ mit $\|z\|_2 = 1$ ein $c \in \mathcal{R}_\ell$ existiert mit

$$\|z - c\|_2 \leq \frac{\eta_1}{\kappa \delta_\ell},$$

wobei $\delta_\ell := \max\left\{\text{diam}(Q_t) \mid t \in \mathcal{T}_\mathcal{I}^\ell\right\}$ sei [3, Kap. 6]. Dann existiert für $\frac{m_t - m_s}{\|m_t - m_s\|_2}$ ein $c \in \mathcal{R}_\ell$, so dass

$$\left\| \frac{m_t - m_s}{\|m_t - m_s\|_2} - c \right\|_2 \leq \frac{\eta_1}{\kappa \delta_\ell} \leq \frac{\eta_1}{\kappa \max\{\text{diam}(Q_t), \text{diam}(Q_s)\}}$$

gilt.

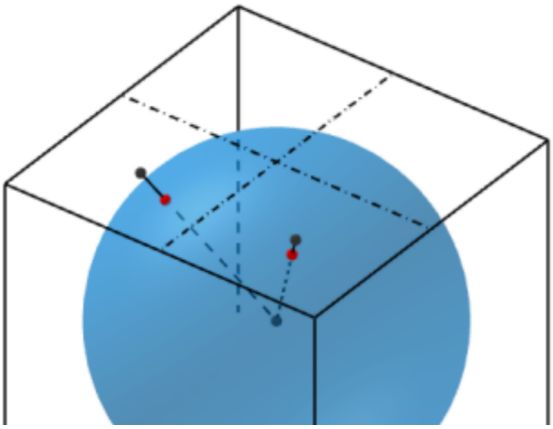


Abbildung 2.15: Projektion vom Würfel auf die Sphäre

Die direkte Konstruktion der Richtungen auf der Oberfläche der Einheitskugel gestaltet sich schwierig, weshalb ein Umweg über den Würfel $[-1, 1]^3$ gemacht wird. Jede Seitenfläche des Würfels wird dazu in Quadrate mit einem Durchmesser von maximal $\frac{2\eta_1}{\kappa \delta_\ell}$ aufgeteilt. Jeder Punkt innerhalb eines solchen Quadrates hat maximal den Abstand $\frac{\eta_1}{\kappa \delta_\ell}$ zum Mittelpunkt, so dass der Mittelpunkt zum Kandidaten für die Richtung wird. Um den Punkt als Richtung zu verwenden, wird er, wie in Abbildung 2.15 gezeigt, auf die Sphäre projiziert. Das Lemma 2.12 garantiert, dass der Abstand zwischen den Punkten durch die Projektion nur verringert werden kann, so dass auch nach der Projektion noch genügend Richtungen vorliegen.
Durch die Art und Weise, wie die praktische Konstruktion der Menge der Richtungen abläuft, kann ihre Mächtigkeit direkt abgeschätzt werden. Da die Seitenflächen des Quaders $[-1, 1]^3$ in Quadrate mit einem Durchmesser von maximal $\frac{2\eta_1}{\kappa \delta_\ell}$ aufgeteilt werden, gibt es

pro Würfelseite höchstens $\left(\lceil \frac{\sqrt{2}\kappa\delta_\ell}{\eta_1} \rceil\right)^2$ viele Richtungen, also insgesamt $6\left(\lceil \frac{\sqrt{2}\kappa\delta_\ell}{\eta_1} \rceil\right)^2$. Da zumindest bis zu der Stufe, bei der die ersten Blätter im Blockbaum auftreten, höchstens so viele Richtungen benötigt werden, wie es Blöcke gibt, kann die tatsächlich verwendete Anzahl an Richtungen in diesen Fällen durch das Minimum aus $6\left(\lceil \frac{\sqrt{2}\kappa\delta_\ell}{\eta_1} \rceil\right)^2$ und $\#\mathcal{T}^\ell_{\mathcal{I}\times\mathcal{I}}$ abgeschätzt werden. Da gerade auf den ersten Stufen viele Richtungen bereitgestellt werden, aber nur wenige Blöcke existieren, steigt die Zahl der tatsächlich benötigten Richtungen zwar mit wachsendem κ und fallendem η_1, aber nicht unbegrenzt.

3 Grundlagen $\mathcal{RH}^2$-Matrizen

Nachdem in den vorherigen Kapiteln das Augenmerk auf den Bedingungen für eine Approximation lag, soll nun endlich die Frage geklärt werden, wie diese eigentlich genau aussieht.

Die Approximation findet nur auf Blöcken $b = (t, s, c) \in \mathcal{L}^+_{\mathcal{I}\times\mathcal{I}}$, die zu Teilmatrizen $A|_{\mathfrak{I}t\times\mathfrak{I}s} \in \mathbb{C}^{\mathfrak{I}t\times\mathfrak{I}s}$ mit $\#^{\mathfrak{I}}t$ Zeilen und $\#^{\mathfrak{I}}s$ Spalten gehören, statt. In vielen Fällen lässt sich das Vorgehen genau wie die algorithmischen Gedankengänge leichter beschreiben, wenn die nicht benötigten Teile der Matrix durch Nullen aufgefüllt werden. Aus diesem Grund verwende die folgenden Formen der Einschränkung und ihre Notationen in dieser Arbeit.

Definition 3.1 (Matrixeinschränkung)
Seien für $M, N \subset \mathbb{N}$ eine Matrix $A \in \mathbb{C}^{M\times N}$, $O \subset M$ sowie $P \subset N$ gegeben, definiere die Einschränkung der Matrix auf die Indexmenge $O \times P$ durch

$$(A|_{O\times P})_{ij} = \begin{cases} a_{i,j} & \text{falls } i \in O, j \in P, \\ 0 & \text{sonst} \end{cases} \qquad \text{für alle } i \in M, j \in N.$$

Um sowohl die Gesamtgröße der Matrix als auch den von null verschiedenen Bereich anzugeben, führe den Raum $\mathbb{C}^{M\times N}_{O\times P}$ ein. Entsprechend bedeutet $A \in \mathbb{C}^{M\times N}_{O\times P}$, dass A $\#M$ Zeilen und $\#N$ Spalten hat, jedoch nur die zu $O \times P$ gehörenden Einträge von null verschieden sein können.
Stammt die Indexmenge von einem Block $(t, s, c) \in \mathcal{T}_{\mathcal{I}\times\mathcal{I}}$, nutze die verkürzte Schreibweise

$$A|_{t\times s} := A|_{\mathfrak{I}t\times\mathfrak{I}s} \in \mathbb{C}^{\mathcal{I}\times\mathcal{I}}_{\mathfrak{I}t\times\mathfrak{I}s}.$$

Soll tatsächlich eine nicht mit Nullen aufgefüllte Teilmatrix betrachtet werden, so notiere dies mit

$$A|_{\star(t\times s)} \in \mathbb{C}^{\mathfrak{I}t\times\mathfrak{I}s}.$$

Verwende die mit Nullen aufgefüllte Einschränkung auch für Vektoren. Entsprechend sei für $N \subset \mathbb{N}$ und eine Teilmenge $P \subset N$ die Einschränkung eines Vektors $x \in \mathbb{C}^N$ durch

$$(x|_P)_j = \begin{cases} x_j & \text{falls } j \in P, \\ 0 & \text{sonst} \end{cases} \qquad \text{für alle } j \in N$$

gegeben. Ebenso nutze auch für Vektoren die verkürzte Schreibweise

$$x|_t = x|_{\mathfrak{I}_t} \in \mathbb{C}^{\mathcal{I}}_{\mathfrak{I}_t}$$

für Indexmengen von einem Cluster $t \in \mathcal{T}_{\mathcal{I}}$.

Bemerkung 13 (Größe der Matrizen in der Praxis): *Das Auffüllen mit Nullen dient rein der Vereinfachung der Schreibweise, so dass die künstlichen Nullblöcke weder in Aufwandsabschätzungen mitgerechnet noch im Computer selbst verwendet werden.*

Im folgenden Kapitel wird definiert, wie eine $\mathcal{RH}^2$-Matrix aufgebaut ist und im darauf folgenden Kapitel erklärt, wie die Approximation mit Hilfe der Polynom-Interpolation berechnet wird. Anschließend wird der Speicheraufwand der Approximation bestimmt und der Ablauf von Algorithmen, die mit $\mathcal{RH}^2$-Matrizen arbeiten, anhand der Matrix-Vektor-Multiplikation beschrieben. Den Schluss dieses Abschnitts bildet ein Kapitel mit Ergebnissen numerischer Experimente zum Aufstellen der $\mathcal{RH}^2$-Matrizen und der Matrix-Vektor-Multiplikation.

3.1 Die $\mathcal{RH}^2$-Matrix

Die Approximation einer Teilmatrix $A|_{t\times s}$ zu einem zulässigen Block $b = (t, s, c) \in \mathcal{L}^+_{\mathcal{I}\times\mathcal{I}}$ soll über ein Produkt von drei Matrizen realisiert werden [6] [3] [10].
Dazu gilt es, möglichst kleine $k_{tc}, k_{sc} \in \mathbb{N}$ zu finden, so dass komplexe Matrizen $V_{tc} \in \mathbb{C}^{\mathcal{I}\times k_{tc}}_{\mathfrak{I}_t\times k_{tc}}$, $S_b \in \mathbb{C}^{k_{tc}\times k_{sc}}$ und $W_{sc} \in \mathbb{C}^{\mathcal{I}\times k_{sc}}_{\mathfrak{I}_s\times k_{sc}}$ existieren mit

$$A|_{t\times s} = V_{tc} S_b W^*_{sc}.$$

Dabei ist $\mathbb{C}^{\mathcal{I}\times k_{sc}}$ so zu verstehen, dass eine Matrix aus diesem Raum $\#\mathcal{I}$ Zeilen mit Indizes $i \in \mathcal{I}$ und k_{sc} Spalten mit Indizes $j \in \underline{k_{sc}|}$ hat.
Bei der obigen Darstellung einer Teilmatrix $A|_{t\times s}$ liegt diese zwar nicht mehr explizit vor, jedoch ist für $k_{tc}, k_{sc} \ll \min\left\{\#^{\mathfrak{I}}t, \#^{\mathfrak{I}}s\right\}$ die Ersparnis im Speicher offensichtlich.
Noch deutlich effizienter ist es, die Matrix V_{tc} für jede mit t und c gebildete Teilmatrix zu verwenden und ebenso W_{sc} für alle Teilmatrizen mit s und c. Selbst wenn k_{tc} beziehungsweise k_{sc} nicht sehr viel kleiner als $\min\left\{\#^{\mathfrak{I}}t, \#^{\mathfrak{I}}s\right\}$ sind, kann durch das mehrfache Verwenden der Matrizen V_{tc} und W_{sc} noch deutlich gespart werden. Diese Idee führt zum Konzept der *Clusterbasis* (vgl. [4, Kap. 3.5]).
Zusätzlich kann die hierarchische Anordnung der Cluster auf das Konzept der Clusterbasen übertragen werden. Betrachte den Cluster $t \in \mathcal{T}_{\mathcal{I}}$. Falls es sich bei t um kein Blattcluster des Clusterbaums handelt, so kann t über seine Kinder $t' \in \mathrm{kind}(t)$ rekonstruiert werden,

da

$$\mathfrak{I}t = \dot{\bigcup_{t' \in \mathrm{kind}(t)}} \mathfrak{I}t'$$

gilt. Genauso soll die Matrix V_{tc} für $c' = \mathrm{r}_t(c)$ über die Matrizen $V_{t'c'}$ seiner Kinder zusammengesetzt werden können. Um dies zu ermöglichen, werden sogenannte *Transfermatrizen* $E_{t'c}$ genutzt, mit denen

$$V_{tc}|_{t' \times k_{tc}} = V_{t'c'} E_{t'c}$$

gilt. Fasse all dies in der Definition der *Clusterbasis* zusammen [3, Def. A.7].

Definition 3.2 (Richtungsabhängige Clusterbasis)
Sei ein richtungsabhängiger Clusterbaum $\mathcal{T}_\mathcal{I}$ zur Indexmenge $\mathcal{I}$ mit der Familie von Richtungsmengen $\mathcal{R}$ gegeben. Bezeichne die Familie von Matrizen $\{V_{tc}\}_{\substack{t \in \mathcal{T}_\mathcal{I} \\ c \in \mathcal{R}_t}}$ als richtungsabhängige Clusterbasis, falls folgende Eigenschaften erfüllt sind

i. $V_{tc} \in \mathbb{C}^{\mathcal{I} \times k_{tc}}_{\mathfrak{I}t \times k_{tc}}$ *für alle* $t \in \mathcal{T}_\mathcal{I}, c \in \mathcal{R}_t$ *mit* $k_{tc} \in \mathbb{N}_0$ *und*

ii. es existiert eine weitere Familie von Matrizen $\{E_{t'c}\}_{\substack{t \in \mathcal{T}_\mathcal{I}, t' \in \mathrm{kind}(t) \\ c \in \mathcal{R}_t}}$ *mit*
$V_{tc}|_{t' \times k_{tc}} = V_{t'c'} E_{t'c}$ *sowie* $E_{t'c} \in \mathbb{C}^{k_{t'c'} \times k_{tc}}$
für alle $t \in \mathcal{T}_\mathcal{I}, t' \in \mathrm{kind}(t), c \in \mathcal{R}_t$ *mit* $c' = \mathrm{r}_t(c)$.

Die Matrizen der Familie $\{E_{t'c}\}_{\substack{t \in \mathcal{T}_\mathcal{I}, t' \in \mathrm{kind}(t) \\ c \in \mathcal{R}_t}}$ *bezeichne als die zur Clusterbasis gehörenden* richtungsabhängigen Transfermatrizen.

Um nicht verwendeten Richtungen gerecht zu werden, ist $k_{tc} = 0$ erlaubt und dank der mit Nullen auffüllenden Definition ist es möglich, die Rekonstruktion der Matrix V_{tc} der Clusterbasis für $t \in \mathcal{T}_\mathcal{I} \setminus \mathcal{L}_\mathcal{I}$ mit $c' = \mathrm{r}_t(c)$ und beliebig vielen Kindern $\#\,\mathrm{kind}(t)$ kompakt mit

$$V_{tc} = \sum_{t' \in \mathrm{kind}(t)} V_{t'c'} E_{t'c}$$

zu formulieren. Die Abbildung 3.1 zeigt das Konzept der Rekonstruktion einer Matrix einer Clusterbasis zu $t \in \mathcal{T}_\mathcal{I} \setminus \mathcal{L}_\mathcal{I}$ und $c \in \mathcal{R}_t$ am Beispiel von zwei Kindern $t_1, t_2 \in \mathrm{kind}(t)$ mit $c' = \mathrm{r}_t(c)$. Nur der schattierte Teil der Matrizen ist dabei von null verschieden und damit von praktischer Relevanz.

Definition 3.3 (Rang Clusterbasis)
Sei zu einem richtungsabhängigen Clusterbaum $\mathcal{T}_\mathcal{I}$ mit Indexmenge $\mathcal{I}$ und Familie von Richtungsmengen $\mathcal{R}$ eine richtungsabhängige Clusterbasis $\{V_{tc}\}_{\substack{t \in \mathcal{T}_\mathcal{I} \\ c \in \mathcal{R}_t}}$ gegeben. Unter dem Rang *einer Matrix $V_{tc} \in \mathbb{C}^{\mathcal{I} \times k_{tc}}_{\mathfrak{I}t \times k_{tc}}$ der Clusterbasis verstehe die Größe k_{tc}.*

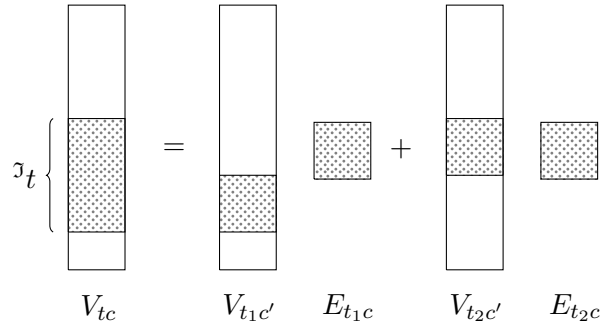


Abbildung 3.1: Darstellung der Matrix V_{tc} einer Clusterbasis über die Matrizen der Kinder t_1, t_2 und Transfermatrizen

Bemerkung 14 (Algebraischer Rang): *Der Rang der Matrizen einer Clusterbasis ist nicht mit dem Objekt des Rangs einer Matrix der linearen Algebra gleichzusetzen, sie können, müssen im Allgemeinen aber nicht übereinstimmen.*

Mit zwei Clusterbasen, wobei eine zu den Spalten- und die andere zu den Zeilenindizes gehört, kann die gesuchte Matrix-Darstellung definiert werden [3, Def. A.8].

Definition 3.4 ($\mathcal{RH}^2$-Matrix)
Seien ein richtungsabhängiger Blockbaum $\mathcal{T}_{\mathcal{I}\times\mathcal{I}}$ mit einer Familie von Richtungsmengen $\mathcal{R}$ sowie zwei zu dem richtungsabhängigen Clusterbaum $\mathcal{T}_{\mathcal{I}}$ gehörende richtungsabhängige Clusterbasen $\{V_{tc}\}_{\substack{t\in\mathcal{T}_{\mathcal{I}}\\ c\in\mathcal{R}_t}}$, $\{W_{sc}\}_{\substack{s\in\mathcal{T}_{\mathcal{I}}\\ c\in\mathcal{R}_s}}$ gegeben. Bezeichne eine Matrix $A \in \mathbb{C}^{\mathcal{I}\times\mathcal{I}}$ als $\mathcal{RH}^2$-Matrix*, falls zwei Familien $\{S_b\}_{b\in\mathcal{L}^+_{\mathcal{I}\times\mathcal{I}}}$ und $\{N_b\}_{b\in\mathcal{L}^-_{\mathcal{I}\times\mathcal{I}}}$ existieren derart, dass die folgenden Eigenschaften erfüllt sind*

i. für alle $b = (t, s, c) \in \mathcal{L}^+_{\mathcal{I}\times\mathcal{I}}$ gilt $A|_{t\times s} = V_{tc} S_b W^*_{sc}$ und

ii. für alle $b = (t, s, c) \in \mathcal{L}^-_{\mathcal{I}\times\mathcal{I}}$ gilt $A|_{\star(t\times s)} = N_b$.

Die Matrizen S_b werden als Kopplungsmatrizen *und N_b als* Nahfeldmatrizen *sowie die Matrizen der Clusterbasen V_{tc} als* Zeilen- *und W_{sc} als* Spaltenclusterbasis *bezeichnet. Bezeichne das Tupel $(\{V_{tc}\}_{\substack{t\in\mathcal{T}_{\mathcal{I}}\\ c\in\mathcal{R}_t}}, \{W_{sc}\}_{\substack{s\in\mathcal{T}_{\mathcal{I}}\\ c\in\mathcal{R}_s}}, \{S_b\}_{b\in\mathcal{L}^+_{\mathcal{I}\times\mathcal{I}}}, \{N_b\}_{b\in\mathcal{L}^-_{\mathcal{I}\times\mathcal{I}}})$ als* $\mathcal{RH}^2$-Matrix-Darstellung *von A.*

Die Abbildung 3.2 zeigt farblich gekennzeichnet, was für eine vollständige $\mathcal{RH}^2$-Matrix-Darstellung benötigt wird. Rote Bereiche entsprechen Teilmatrizen aus dem Nahfeld, sie müssen vollbesetzt gespeichert werden. Grüne Quader symbolisieren die Kopplungsmatrizen, sie sind zum Teil erheblich kleiner als die ursprünglichen Teilmatrizen. In blau sind die Bestandteile der Clusterbasen markiert. Die äußeren hellblauen Quader repräsentieren die Matrizen auf der höchsten Stufe, die links beziehungsweise oberhalb liegenden klei-

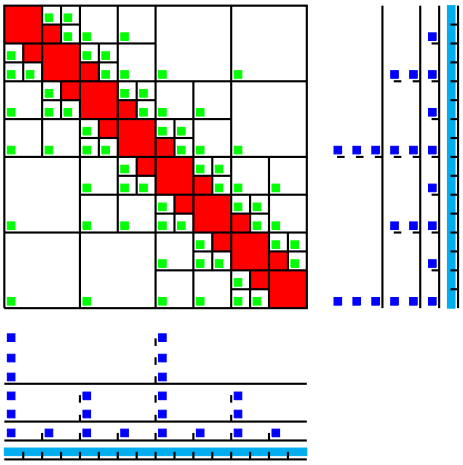


Abbildung 3.2: Speichermuster einer $\mathcal{RH}^2$-Matrix-Darstellung

nen dunkelblauen Quadrate stehen für Transfermatrizen. Potentiell kann eine Clusterbasis zum Cluster $t \in \mathcal{T}_\mathcal{I}$ für jede Richtung auf Stufe $\mathrm{stufe}(t)$ eine Matrix beinhalten, was durch mehrfaches Darstellen der blauen Quader auf einer Stufe in der Abbildung angedeutet ist.

Definition 3.5 (Rang einer $\mathcal{RH}^2$-Matrix)
Unter dem Rang k einer $\mathcal{RH}^2$-Matrix A zum richtungsabhängigen Blockbaum $\mathcal{T}_{\mathcal{I}\times\mathcal{I}}$ mit der Familie an Richtungsmengen $\mathcal{R}$ verstehe den maximalen Rang der zugehörigen richtungsabhängigen Clusterbasen, entsprechend gilt

$$k \geq k_{tc}, k_{sc} \qquad \text{für alle } t \in \mathcal{T}_\mathcal{I}, s \in \mathcal{T}_\mathcal{I} \text{ und für alle } c \in \mathcal{R}_t = \mathcal{R}_s.$$

Bemerkung 15 (Unterschied zu klassischen $\mathcal{H}^2$-Matrizen): *Im Unterschied zu einer klassischen $\mathcal{H}^2$-Matrix können Matrizen der Clusterbasen und Transfermatrizen zu einem Cluster $t \in \mathcal{T}_\mathcal{I}$ in allen möglichen Richtungen $c \in \mathcal{R}_t$ auftreten. Die gesamte richtungsabhängige Clusterbasis ist damit tendenziell deutlich größer als eine entsprechende Clusterbasis ohne Richtungen (Bei dieser Überlegung sei derselbe Blockbaum für beide Matrizen zugrunde gelegt!). Die Anzahl der Kopplungsmatrizen bleibt unangetastet, da ein zulässiger Block nur in Zusammenhang mit einer einzigen Richtung $c \in \mathcal{R}_t$ zulässig ist.*

In der Praxis wird nur mit einer Approximation der Matrix A in Form einer $\mathcal{RH}^2$-Matrix gearbeitet und nicht mit einer exakten Matrix-Darstellung. Um diesem Fakt gerecht zu werden, bezeichne eine $\mathcal{RH}^2$-Matrix $\widetilde{A}$ mit $\mathcal{RH}^2$-Matrix-Darstellung
$(\{V_{tc}\}_{\substack{t\in\mathcal{T}_\mathcal{I}\\ c\in\mathcal{R}_t}}, \{W_{sc}\}_{\substack{s\in\mathcal{T}_\mathcal{I}\\ c\in\mathcal{R}_s}}, \{S_b\}_{b\in\mathcal{L}^+_{\mathcal{I}\times\mathcal{I}}}, \{N_b\}_{b\in\mathcal{L}^-_{\mathcal{I}\times\mathcal{I}}})$ zu einem richtungsabhängigen Blockbaum $\mathcal{T}_{\mathcal{I}\times\mathcal{I}}$ mit einer Familie an Richtungsmengen $\mathcal{R}$ als *$\mathcal{RH}^2$-Matrix-Approximation* von $A \in \mathbb{R}^{\mathcal{I}\times\mathcal{I}}$, falls

i. für alle $b = (t, s, c) \in \mathcal{L}^-_{\mathcal{I}\times\mathcal{I}}$ gilt $A|_{\star(t\times s)} = N_b$ und

ii. für alle $b = (t, s, c) \in \mathcal{L}^+_{\mathcal{I}\times\mathcal{I}}$ mit einem hinreichend kleinen Fehler $A|_{t\times s} \approx V_{tc}S_bW_{sc}$ erfüllt ist.

Es bleibt die nicht triviale Frage, wie eine entsprechende Matrix-Approximation nun genau mit Hilfe der Polynom-Interpolation gewonnen werden kann.

3.1.1 Approximation via Interpolation

Eine Möglichkeit, eine $\mathcal{RH}^2$-Matrix-Approximation zu berechnen, besteht in der Verwendung der Polynom-Interpolation [3]. Der Definitionsbereich der Kernfunktionen ist $\Gamma\times\Gamma \subset \mathbb{R}^3 \times \mathbb{R}^3$. Die resultierende sechsdimensionale Tensorinterpolation nutzt in jeder Richtung die Ordnung $m \in \mathbb{N}_0$ und wird für die Berechnung der Clusterbasen und Kopplungsmatrizen ebenfalls als Kombination zweier dreidimensionaler Interpolationen geschrieben. Entsprechend sei im Folgenden $\widehat{M} = M^3$ und $\widehat{\mu} \in \widehat{M}$ ein dreidimensionaler Multiindex.
Für einen zulässigen richtungsabhängigen Block $b = (t, s, c) \in \mathcal{T}_{\mathcal{I}\times\mathcal{I}}$ nutze zwei Familien von Interpolationspunkten $\left\{x_{\widehat{\mu}}\right\}_{\widehat{\mu}\in\widehat{M}} \in Q_t$ und $\{y_{\widehat{\nu}}\}_{\widehat{\nu}\in\widehat{M}} \in Q_s$, welche in jeder Koordinatenrichtung durch Transformationen aus denen auf $[-1, 1]$ hervorgehen. Entsprechend verwende zwei Familien von Lagrange-Polynomen $\left\{\ell_{t,\widehat{\mu}}\right\}_{\widehat{\mu}\in\widehat{M}}, \left\{\ell_{s,\widehat{\nu}}\right\}_{\widehat{\nu}\in\widehat{M}}$.
Da für die Interpolation die Modifikation der Kernfunktion mit Hilfe ebener Wellen notwendig ist, bietet es sich an, den Interpolationsoperator (siehe Definition 2.6) mit

$$\mathfrak{I}_{Q_t}[u](x) = \sum_{\widehat{\mu}\in\widehat{M}} u(x_{\widehat{\mu}})\ell_{t,\widehat{\mu}}(x) \qquad \text{für alle } u \in C(Q_t), x \in Q_t$$

ebenfalls zu modifizieren und einen richtungsangepassten Interpolationsoperator zu definieren [3, Kap. 2.2.1]

$$\begin{aligned} \mathfrak{I}^c_{Q_t}[u](x) &:= e^{\mathrm{i}\kappa\langle x,c\rangle_2}\mathfrak{I}_{Q_t}[e^{-\mathrm{i}\kappa\langle x,c\rangle_2}u](x) && \text{für alle } u \in C(Q_t), x \in Q_t,\\ \mathfrak{I}^{-c}_{Q_s}[u](y) &:= e^{\mathrm{i}\kappa\langle y,-c\rangle_2}\mathfrak{I}_{Q_s}[e^{-\mathrm{i}\kappa\langle y,-c\rangle_2}u](y) && \text{für alle } u \in C(Q_s), y \in Q_s. \end{aligned} \tag{3.1.1}$$

Damit ergibt sich für alle $u \in C(Q_t \times Q_s)$ mit $(x, y) \in Q_t \times Q_s$

$$\mathfrak{I}^c_{Q_t\times Q_s}[u](x, y) := \left(\mathfrak{I}^c_{Q_t} \otimes \mathfrak{I}^{-c}_{Q_s}\right)[u](x, y) = e^{\mathrm{i}\kappa\langle x-y,c\rangle_2}\mathfrak{I}_{Q_t\times Q_s}[e^{-\mathrm{i}\kappa\langle x-y,c\rangle_2}u](x, y).$$

Auf diese Weise kann die Interpolation leicht mit der modifizierten Kernfunktion des Einfachschichtoperators (2.2.6)

$$g_{ec} = \frac{e^{\mathrm{i}\kappa(\|x-y\|_2-\langle x-y,c\rangle_2)}}{4\pi\|x-y\|_2}$$

als

$$\begin{aligned} \mathfrak{I}^c_{Q_t\times Q_s}[g_e](x, y) &= e^{\mathrm{i}\kappa\langle x-y,c\rangle_2}\mathfrak{I}_{Q_t\times Q_s}[e^{-\mathrm{i}\kappa\langle x-y,c\rangle_2}g_e](x, y)\\ &= e^{\mathrm{i}\kappa\langle x-y,c\rangle_2}\mathfrak{I}_{Q_t\times Q_s}[g_{ec}](x, y) \end{aligned}$$

geschrieben werden. Die beim Modifizieren der Kernfunktion zusätzlich hinzugekommenen zwei Exponentialterme $e^{\mathrm{i}\kappa\langle x,c\rangle_2}$ und $e^{\mathrm{i}\kappa\langle y,-c\rangle_2}$ können für alle $c \in \mathcal{R}_t$ und $\widehat{\mu}, \widehat{\nu} \in \widehat{M}$ mit den Lagrange-Polynomen zusammengefasst werden

$$\begin{aligned} \ell_{tc,\widehat{\mu}}(x) &:= \ell_{t,\widehat{\mu}}(x) e^{\mathrm{i}\kappa\langle x,c\rangle_2} && \text{für } x \in Q_t \quad \text{beziehungsweise} \\ \ell_{sc,\widehat{\nu}}(x) &:= \ell_{s,\widehat{\nu}}(y) e^{\mathrm{i}\kappa\langle y,c\rangle_2} && \text{für } y \in Q_s, \end{aligned} \tag{3.1.2}$$

so dass die modifizierten Lagrange-Polynome ebenfalls einen richtungsabhängigen Anteil in Form einer ebenen Welle aufweisen [3, Kap. 2.2.1]. Das Lagrange-Polynom für den Spaltencluster s wird später komplex konjugiert, weshalb das modifizierte Lagrange-Polynome ebenfalls mit c im zweitem Argument des Skalarprodukts definiert wurde. Mit dem angepassten Interpolationsoperator kann es an die Bestimmung der Matrixeinträge gehen.
Betrachte zunächst die Kernfunktion g_e und die Matrix A_e des Einfachschichtoperators. Die Kernfunktion lässt sich auf dem Block $b = (t, s, c) \in \mathcal{L}^+_{\mathcal{I}\times\mathcal{I}}$ mit den modifizierten Lagrange-Polynomen (3.1.2) für alle $x \in Q_t, y \in Q_s$ wie folgt interpolieren

$$g_e(x,y) \approx \mathfrak{I}^c_{Q_t\times Q_s}[g_e](x,y) = \sum_{\widehat{\mu}\in\widehat{M}} \sum_{\widehat{\nu}\in\widehat{M}} g_{ec}(x_{\widehat{\mu}}, y_{\widehat{\nu}}) \ell_{tc,\widehat{\mu}}(x) \overline{\ell_{sc,\widehat{\nu}}(y)}.$$

Die dreigliedrige Produktform der Approximation der Teilmatrix $A_e|_{t\times s}$ lässt sich in dieser Schreibweise schon erahnen.
Für den mit Hilfe der Integralformulierung (siehe Kapitel 1.3.3) aufgestellten Matrixeintrag $(A_e)_{ij}$ mit $i \in {}^{\mathfrak{I}}t, j \in {}^{\mathfrak{I}}s$ und den reellwertigen Basisfunktionen ϕ_i, ϕ_j gilt [3, G. 2.10]

$$\begin{aligned} (A_e)_{ij} &= \int_\Gamma \phi_i(x) \int_\Gamma g_e(x,y)\phi_j(y)\,\mathrm{d}y\,\mathrm{d}x \\ &\approx \int_\Gamma \phi_i(x) \int_\Gamma \sum_{\widehat{\mu}\in\widehat{M}} \sum_{\widehat{\nu}\in\widehat{M}} g_{ec}(x_{\widehat{\mu}}, y_{\widehat{\nu}}) \ell_{tc,\widehat{\mu}}(x) \overline{\ell_{sc,\widehat{\nu}}(y)} \phi_j(y)\,\mathrm{d}y\,\mathrm{d}x \\ &= \sum_{\widehat{\mu}\in\widehat{M}} \sum_{\widehat{\nu}\in\widehat{M}} \int_\Gamma \phi_i(x)\ell_{tc,\widehat{\mu}}(x)\,\mathrm{d}x\; g_{ec}(x_{\widehat{\mu}}, y_{\widehat{\nu}}) \int_\Gamma \overline{\ell_{sc,\widehat{\nu}}(y)\phi_j(y)}\,\mathrm{d}y =: (\widetilde{A}_e)_{ij}. \end{aligned}$$

Setze die Nichtnulleinträge der Matrizen für die Clusterbasen und die Kopplungsmatrizen, entsprechend der in der Gleichung schon angedeuteten drei Faktoren, mit

$$\begin{aligned} (S_b)_{\widehat{\mu}\widehat{\nu}} &= g_{ec}(x_{\widehat{\mu}}, y_{\widehat{\nu}}) && \text{für alle } b \in \mathcal{L}^+_{\mathcal{I}\times\mathcal{I}}, \widehat{\mu} \in \widehat{M}, \widehat{\nu} \in \widehat{M}, \\ (V_{tc})_{i\widehat{\mu}} &= \int_\Gamma \phi_i(x)\ell_{tc,\widehat{\mu}}(x)\,\mathrm{d}x && \text{für alle } t \in \mathcal{T}_\mathcal{I}, i \in {}^{\mathfrak{I}}t, c \in \mathcal{R}_t, \widehat{\mu} \in \widehat{M}, \\ (W_{sc})_{j\widehat{\nu}} &= \int_\Gamma \phi_j(y)\ell_{sc,\widehat{\nu}}(y)\,\mathrm{d}y && \text{für alle } s \in \mathcal{T}_\mathcal{I}, j \in {}^{\mathfrak{I}}s, c \in \mathcal{R}_s, \widehat{\nu} \in \widehat{M}, \end{aligned}$$

womit die gewünschte Faktorisierung folgt

$$A_e|_{t\times s} \approx V_{tc} S_b W^*_{sc}.$$

In der Definition der Clusterbasis wurde zusätzlich gefordert, dass sich die Matrix $V_{t^+c^+}$ für t^+ mit $\operatorname{kind}(t^+) \neq \emptyset$ und $\mathrm{r}_{t^+}(c^+) = c$ aus den Matrizen $\{V_{tc}\}_{t\in\operatorname{kind}(t^+)}$ mit Hilfe von Transfermatrizen zusammensetzen lässt. Dies kann über eine Reinterpolation der Lagrange-Polynome des Elternclusters geschehen [3, Kap. 2.2.2]. Seien dazu die Stützpunkte des Kinds t mit $\left\{x_{\widehat{\mu}}\right\}_{\widehat{\mu}\in\widehat{M}}$ und die des Elternclusters t^+ mit $\left\{x^+_{\widehat{\mu}^+}\right\}_{\widehat{\mu}^+\in\widehat{M}}$ gegeben. Sei $c^+ \in \mathcal{R}_{t^+}$ die Richtung im Elterncluster und $c = \mathrm{r}_{t^+}(c^+)$ seine Approximation im Kind, dann kann für $\widehat{\mu}^+ \in \widehat{M}$ das Lagrange-Polynom $\ell_{t^+c^+,\widehat{\mu}^+}$ des Elternclusters für alle $x \in Q_t$ mit

$$\begin{aligned}
\ell_{t^+c^+,\widehat{\mu}^+}(x) &= e^{\mathrm{i}\kappa\langle x,c^+\rangle_2}\ell_{t^+,\widehat{\mu}^+}(x) = e^{\mathrm{i}\kappa\langle x,c\rangle_2}e^{\mathrm{i}\kappa\langle x,c^+-c\rangle_2}\ell_{t^+,\widehat{\mu}^+}(x)\\
&\approx e^{\mathrm{i}\kappa\langle x,c\rangle_2}\sum_{\widehat{\mu}\in\widehat{M}} e^{\mathrm{i}\kappa\langle x_{\widehat{\mu}},c^+-c\rangle_2}\ell_{t^+,\widehat{\mu}^+}(x_{\widehat{\mu}})\ell_{t,\widehat{\mu}}(x)\\
&= \sum_{\widehat{\mu}\in\widehat{M}} \underbrace{e^{\mathrm{i}\kappa\langle x_{\widehat{\mu}},c^+-c\rangle_2}\ell_{t^+,\widehat{\mu}^+}(x_{\widehat{\mu}})}_{=:(E_{tc^+})_{\widehat{\mu}\widehat{\mu}^+}} \underbrace{e^{\mathrm{i}\kappa\langle x,c\rangle_2}\ell_{t,\widehat{\mu}}(x)}_{=\ell_{tc,\widehat{\mu}}(x)}
\end{aligned}$$

approximiert werden, was eine Möglichkeit zur Schachtelung bietet.
Nachdem geklärt ist, wie die Matrixeinträge beim Einfachschichtoperator gewonnen werden, soll nun die Kernfunktion des Doppelschichtoperators betrachtet werden. Beim Berechnen des Doppelschichtoperators wird die Normalenableitung erst nach dem Interpolieren gebildet. Für einen Block $(t, s, c) \in \mathcal{L}^+_{\mathcal{I}\times\mathcal{I}}$ und $x \in Q_t, y \in Q_s$ gilt

$$\begin{aligned}
\tfrac{\partial}{\partial n(y)}g_e(x,y) &\approx \tfrac{\partial}{\partial n(y)}\mathfrak{I}^c_{Q_t\times Q_s}[g_e](x,y)\\
&= \tfrac{\partial}{\partial n(y)}\sum_{\widehat{\mu}\in\widehat{M}}\sum_{\widehat{\nu}\in\widehat{M}} g_{ec}(x_{\widehat{\mu}},y_{\widehat{\nu}})\ell_{tc,\widehat{\mu}}(x)\overline{\ell_{sc,\widehat{\nu}}(y)}\\
&= \sum_{\widehat{\mu}\in\widehat{M}}\sum_{\widehat{\nu}\in\widehat{M}} g_{ec}(x_{\widehat{\mu}},y_{\widehat{\nu}})\ell_{tc,\widehat{\mu}}(x)\tfrac{\partial}{\partial n(y)}\overline{\ell_{sc,\widehat{\nu}}(y)}.
\end{aligned}$$

Damit wirkt sich die Differentiation nur auf die Lagrange-Polynome der Spalten aus. Die Nichtnulleinträge der Approximation der Matrix des Doppelschichtoperators $(\widetilde{A}_d)_{ij}$ mit reellen Basisfunktionen ϕ_i, ϕ_j ergeben sich durch die drei Matrizen mit

$$\begin{aligned}
(S_b)_{\widehat{\mu}\widehat{\nu}} &= g_{ec}(x_{\widehat{\mu}},y_{\widehat{\nu}}) && \text{für alle } b \in \mathcal{L}^+_{\mathcal{I}\times\mathcal{I}}, \widehat{\mu} \in \widehat{M}, \widehat{\nu} \in \widehat{M},\\
(V_{tc})_{i\widehat{\mu}} &= \int_\Gamma \phi_i(x)\ell_{tc,\widehat{\mu}}(x)\,\mathrm{d}x && \text{für alle } t \in \mathcal{T}_\mathcal{I}, i \in {}^\mathfrak{I}t, c \in \mathcal{R}_t, \widehat{\mu} \in \widehat{M},\\
(W_{sc})_{j\widehat{\nu}} &= \int_\Gamma \phi_j(y)\tfrac{\partial}{\partial n(y)}\ell_{sc,\widehat{\nu}}(y)\,\mathrm{d}y && \text{für alle } s \in \mathcal{T}_\mathcal{I}, j \in {}^\mathfrak{I}s, c \in \mathcal{R}_s, \widehat{\nu} \in \widehat{M}.
\end{aligned}$$

Die Schachtelungseigenschaft lässt sich auch in diesem Fall mit Hilfe von Reinterpolation sicherstellen. Seien ein $s^+ \in \mathcal{T}_\mathcal{I}$ mit $s \in \operatorname{kind}(s^+)$ und der Richtung $c^+ \in \mathcal{R}_{s^+}$ gegeben,

für $c = \mathrm{r}_{s^+}(c^+)$ und alle $\widehat{\nu}^+ \in \widehat{M}$ und $y \in Q_s$ folgt

$$\begin{aligned}\frac{\partial}{\partial n(y)}\ell_{s^+c^+,\widehat{\nu}^+}(y) &\approx \frac{\partial}{\partial n(y)}e^{\mathrm{i}\kappa\langle y,c\rangle_2}\sum_{\widehat{\nu}\in\widehat{M}} e^{\mathrm{i}\kappa\langle y_{\widehat{\nu}},c^+-c\rangle_2}\ell_{s^+,\widehat{\nu}^+}(y_{\widehat{\nu}})\ell_{s,\widehat{\nu}}(y)\\ &= \sum_{\widehat{\nu}\in\widehat{M}} \underbrace{e^{\mathrm{i}\kappa\langle y_{\widehat{\nu}},c^+-c\rangle_2}\ell_{s^+,\widehat{\nu}^+}(y_{\widehat{\nu}})}_{=:(E_{sc^+})_{\widehat{\nu}\widehat{\nu}+}}\frac{\partial}{\partial n(y)}\ell_{sc,\widehat{\nu}}(y).\end{aligned}$$

Entsprechend unterscheiden sich die Transfermatrizen im Fall des Doppelschichtoperators nicht von denen des Einfachschichtoperators, der Unterschied besteht in den Matrizen der Blattcluster.

Bemerkung 16 (Rang): *Beim Ansatz der Interpolation haben alle auftretenden Matrizen der Clusterbasis denselben Rang k, der durch $k = |\widehat{M}| = (m+1)^3$ gegeben ist. Das heißt, für alle nicht leeren Matrizen V_{tc} der Clusterbasis mit $t \in \mathcal{T}_\mathcal{I}$ und $c \in \mathcal{R}_t$ gilt $k_{tc} = k$.*

3.1.2 Aufwand

Ziel dieses Abschnitts ist es, Abschätzungen für den Speicheraufwand einer $\mathcal{RH}^2$-Matrix-Approximation herzuleiten, dazu müssen jedoch zunächst noch Aussagen zu der Anzahl der vorhandenen Blöcke und Richtungen getroffen werden.

Alle Cluster $t \in \mathcal{T}_\mathcal{I}^\ell$ nutzen dieselbe Mengen an Richtungen $\mathcal{R}_\ell$, jedoch taucht meist im Zusammenhang mit einem festen Cluster t nur ein Teil der Richtungen auf. Die zu den ungenutzten Richtungen gehörenden Matrizen der Clusterbasis sind leer und sollen folglich in den Abschätzungen der Algorithmen nicht auftauchen. Um der Diskrepanz zwischen den möglichen und verwendeten Richtungen gerecht zu werden, führe die Menge der *effektiven Richtungen* $\mathcal{R}_t^{eff}$ ein[i].

Definition 3.6 (Effektive Richtungen)
Die Menge der effektiven Richtungen für ein Cluster $t \in \mathcal{T}_\mathcal{I}$ ist definiert durch

$$\mathcal{R}_t^{eff} := \{c \in \mathcal{R}_t \mid \exists (t,s,c) \in \mathcal{T}_{\mathcal{I}\times\mathcal{I}}\}.$$

Die effektiven Richtungen $\mathcal{R}_t^{eff}$ tauchen auch bei der Beschränkung der Anzahl der Spaltenclusterpartner $\mathrm{row}(t)$ (siehe Definition 2.30) auf

$$\#\,\mathrm{row}(t) = \#\bigcup_{c\in\mathcal{R}_t^{eff}} \mathrm{row}_c(t) \leq \#\mathcal{R}_t^{eff}\max\{\#\,\mathrm{row}_c(t) \mid c \in \mathcal{R}_t\}. \tag{3.1.3}$$

[i] In dem Kapitel zu den numerischen Experimenten des Aufwands befindet sich eine Tabelle 3.1, die den Unterschied zwischen $\mathcal{R}_t$ und $\mathcal{R}_t^{eff}$ zeigt.

Eine gute Schranke für die Mächtigkeit von $\mathcal{R}_t^{eff}$ ist dabei nicht so einfach zu finden. Natürlich gilt immer $\#\mathcal{R}_t^{eff} \leq \#\mathcal{R}_t$, dies dürfte in den meisten Fällen jedoch eine deutlich zu grobe Abschätzung sein. Wie bei der Konstruktion der Richtungen (siehe Ende Kapitel 2.5) schon erwähnt, stehen auf den niedrigen Stufen oftmals sehr viel mehr Richtungen zur Verfügung als es überhaupt Blöcke gibt. Folglich gilt in diesen Fällen $\#\mathcal{R}_t^{eff} \ll \#\mathcal{R}_t$.
Aufgrund der Konstruktion von $\#\mathcal{R}_t$ kann jedoch davon ausgegangen werden, dass für eine feste Geometrie eine Konstante $\mathcal{C}_\mathcal{D} \in \mathbb{R}_{>0}$ existiert, so dass für jede Stufe

$$\#\mathcal{R}_t^{eff} \leq 1 + \kappa^2 \mathcal{C}_\mathcal{D} \operatorname{diam}_{\max}(\ell)^2 \qquad \text{für alle } \ell \in \underline{p_{\mathcal{I}]}}\, 0, t \in \mathcal{T}_\mathcal{I}^\ell \tag{3.1.4}$$

mit den maximalen und minimalen Durchmessern von Clustern auf Stufe $\ell \in \underline{p_{\mathcal{I}]}}\, 0$

$$\operatorname{diam}_{\max}(\ell) := \max\left\{\operatorname{diam}(Q_t) \mid t \in \mathcal{T}_\mathcal{I}^\ell\right\} > 0,$$
$$\operatorname{diam}_{\min}(\ell) := \min\left\{\operatorname{diam}(Q_t) \mid t \in \mathcal{T}_\mathcal{I}^\ell\right\} > 0$$

erfüllt ist [10]. Die Eins in dieser Abschätzung ist nötig, da auf jeder Stufe mindestens eine Richtung vorliegt.
Um eine Schranke für die Anzahl von Blöcken zu erhalten, sind deutlich tiefer gehende Überlegungen notwendig. Dabei erweist sich der richtungsabhängige Blockbaum als unhandlich in Komplexitätsabschätzungen, während deutlich mehr Aussagen zu richtungsabhängigen Clusterbäumen möglich sind. Um den Blockbaum auf den Clusterbaum zurückzuführen, kann das Konzept der Schwachbesetztheit verwendet werden. Die Idee wurde erstmals von Lars Grasedyck [27, S. 67] auf $\mathcal{H}$-Matrizen übertragen und verknüpft die Eigenschaft von Matrizen, schwach besetzt zu sein, also in jeder Zeile und Spalte nur eine begrenzte und im Vergleich zur Dimension kleine Anzahl von Einträgen ungleich null zu besitzen, mit Blockbäumen. Im Kontext der Blockbäume ist die Eigenschaft dann so zu verstehen, dass die Anzahl der Blöcke, die mit einem festen Spalten- oder Zeilencluster gebildet werden, durch eine Konstante beschränkt ist.
Um eine Aussage darüber zu gewinnen, ob eine Schwachbesetztheitskonstante existiert und wie sie aussieht, wird zunächst die Anzahl der möglichen Clusterpartner eines konkreten Clusters s untersucht. Doch auch bei diesem Ansatz gestaltet sich das Bestimmen einer Abschätzung über einen direkten Zugang schwierig. Deutlich einfacher ist es, unzulässige Zeilen- oder Spaltenclusterpartner des Elternclusters zu betrachten und auf diese Weise die Anzahl der möglichen Zeilen- oder Spaltenclusterpartner des Kindes abzuschätzen. Dies ist möglich, da bei der hier genutzten Konstruktion des richtungsabhängigen Blockbaums nur unzulässige Blöcke Kinder haben können.
Die Kernidee für die Abschätzung ist es, ein Volumenargument zu verwenden, bei dem die maximale Zahl der Cluster, die unter Berücksichtigung der Überlappung unzulässige Zeilen- oder Spaltenclusterpartner sein können, bestimmt wird. Für einen Cluster s und eine Richtung c legt der Aufbau der Approximation nahe, einen Kegel als maximale Volumenmenge für mögliche unzulässige Zeilen- oder Spaltenclusterpartner t zu verwenden.

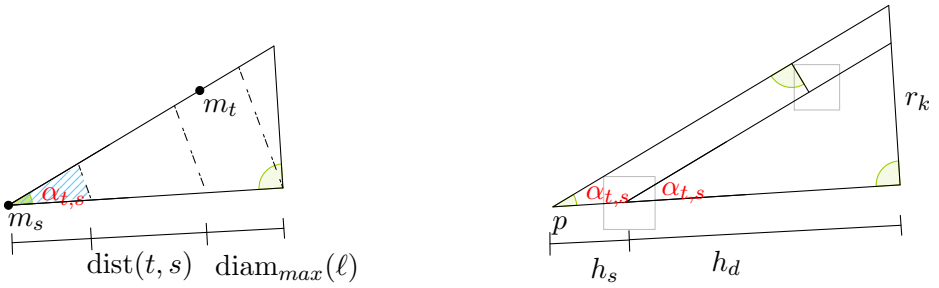
Die unzulässigen Clusterpartner t selbst werden von Kugeln umschlossen. Dabei muss jedoch bedacht werden, dass die Cluster nur auf der Oberfläche der Geometrie Γ liegen, die betrachteten Volumen sollten entsprechend eingeschränkt werden. Die Anzahl der maximal möglichen unzulässigen Clusterpartner ist dann durch den Quotienten aus dem Kegel- und Kugelvolumen beschränkt. Falls die Menge der Richtungen nur die Null enthält, wird mit einer Kugel um s statt mit einem Kegel gearbeitet.

Für die bessere Anschauung erfolgt die Entwicklung des Kegels zum Cluster s zunächst am Beispiel eines unzulässigen Clusterpartners t. Anschließend wird dieser konkrete Kegel mit Hilfe der Zulässigkeitsbedingungen, die alle unzulässigen Clusterpartner nicht erfüllen, verallgemeinert und von der konkreten Richtung unabhängig gemacht, so dass schließlich alle unzulässigen Clusterpartner entlang einer Richtung enthalten sind.
Sei t im Folgenden ein fester Clusterpartner, so dass (t, s, c) ein unzulässiger richtungsabhängiger Block ist und sei $\ell = \mathrm{stufe}(s)$ seine Stufe im Baum. Die Ausgangssituation ist in Abbildung 3.3 (a) dargestellt. Die Richtung c und die normierte Differenz der Mittel-

(a) Positionen der Cluster (b) Ausgangsdreieck

Abbildung 3.3: Ausgangslage

punkte $\frac{m_t - m_s}{\|m_t - m_s\|_2}$ bilden ein gleichschenkliges Ausgangsdreieck siehe Abb. 3.3 (b), dabei entspricht der eingeschlossene Winkel $\alpha_{t,s}$ dem halben Öffnungswinkel des Kegels.
Im nächsten Schritt wird das Ausgangsdreieck gestreckt, so dass die Schenkel des Dreiecks garantiert den Rand des überdeckenden Quaders von s erreichen. Da die Schenkel c und $\frac{m_t - m_s}{\|m_t - m_s\|_2}$ auf eine Länge von eins normiert sind, reicht hierzu die Verlängerung der Schenkel auf eine Länge von $\frac{\mathrm{diam}_{\max}(\ell)}{2}$ aus. Das Ergebnis dieser Streckung ist das gestrichelte Dreieck in Abbildung 3.4 (a).
In einem zweiten Schritt wird das Dreieck noch einmal erweitert, so dass der in Richtung des Clusters t zeigende Schenkel des Dreiecks, den Cluster t mindestens vollständig passiert. Dazu werden die Schenkel additiv um die Distanz der beiden Cluster $\mathrm{dist}(Q_t, Q_s)$ und $\mathrm{diam}_{\max}(\ell)$ verlängert, in Abbildung 3.4 (a) ist $\mathrm{dist}(Q_t, Q_s)$ durch die Kurzform $\mathrm{dist}(t, s)$ gekennzeichnet.
Zur Bildung eines leicht berechenbaren Kegels ist ein rechtwinkliges Dreieck nötig. Dafür soll die Erweiterung der Richtung c als Höhe des Kegels genutzt werden, weshalb der zwei-



(a) Rechtwinkliges Dreieck

(b) Endgültige Dreieckserweiterung

Abbildung 3.4: Erweiterungen des Dreiecks

te Schenkel (dieser verbindet die beiden Mittelpunkte der Cluster s und t miteinander) zur Hypotenuse verlängert wird. Das so entstehende rechtwinklige Dreieck ist in Abbildung 3.4 (a) zu sehen. In diesem Dreieck ist der überdeckende Quader des Clusters t jedoch immer noch nicht vollständig enthalten, denn die Hypotenuse des Dreiecks läuft durch den Mittelpunkt m_t des Clusters. Um den Cluster t vollständig zu überdecken, wird die Hypotenuse parallel um den Abstand $\frac{\mathrm{diam}_{\max}(\ell)}{2}$ verschoben, so dass ein größeres rechtwinkliges Dreieck mit neuer Spitze p statt m_s entsteht. Der Radius der Grundfläche des Kegels soll mit r_k bezeichnet werden. Das Ergebnis der Parallelverschiebung ist in Abbildung 3.4 (b) zu sehen und bildet das Grundgerüst des Kegels. In der Abbildung 3.4 (b) sind zudem in hellgrau die Cluster s und t eingezeichnet.
Noch hängt der Kegel vom festen Clusterpartner t ab, betrachtet werden sollen jedoch alle unzulässige Blöcke und nicht nur ein spezielles Clusterpaar. Um die Abhängigkeit von dem Clusterpartner t zu eliminieren, können die Zulässigkeitsbedingungen (2.3.2c), (2.3.2b) genutzt werden. Denn wenn die Standardzulässigkeitsbedingung

$$\max\{\mathrm{diam}(Q_t), \mathrm{diam}(Q_s)\} \leq \eta_2 \,\mathrm{dist}(Q_t, Q_s)$$

beziehungsweise die parabolische Zulässigkeitsbedingung

$$\kappa \max\left\{\mathrm{diam}^2(Q_t), \mathrm{diam}^2(Q_s)\right\} \leq \eta_2 \,\mathrm{dist}(Q_t, Q_s)$$

nicht erfüllt sind, gelten entsprechend

$$\tfrac{1}{\eta_2} \max\{\mathrm{diam}(Q_t), \mathrm{diam}(Q_s)\} > \mathrm{dist}(Q_t, Q_s)$$

beziehungsweise

$$\tfrac{\kappa}{\eta_2} \max\left\{\mathrm{diam}^2(Q_t), \mathrm{diam}^2(Q_s)\right\} > \mathrm{dist}(Q_t, Q_s),$$

mit denen sich der Anteil h_d der Höhe

$$\tfrac{1}{2}\,\mathrm{diam}_{\max}(\ell) + \mathrm{dist}(Q_t, Q_s) + \mathrm{diam}_{\max}(\ell)$$

unabhängig vom Cluster t abschätzen lässt. Definiere dazu den von der Zulässigkeitsbedingung abhängigen Anteil $r_{zb}(\ell)$ mit

$$r_{zb}(\ell) := \begin{cases} \frac{3}{2} + \frac{1}{\eta_2} & \text{falls } \kappa \operatorname{diam}_{\max}(\ell) \leq 1, \\ \frac{3}{2} + \frac{\operatorname{diam}_{\max}(\ell)\,\kappa}{\eta_2} & \text{sonst,} \end{cases}$$

womit der vom konkreten Clusterpaar unabhängige Anteil h_d der Höhe durch

$$h_d = \operatorname{diam}_{\max}(\ell)\, r_{zb}(\ell)$$

gegeben ist.
Auch der Öffnungswinkel des Kegels hängt aktuell noch von dem konkreten Clusterpartner t ab und ist durch $2\alpha_{t,s}$ gegeben. Da der Kegel jedoch alle unzulässigen Clusterpartner entlang der Richtung c enthalten soll, muss der Öffnungswinkel vergrößert werden. Die maximalen Öffnungswinkel der Kegel werden durch die Lage der einzelnen Richtungen auf der Stufe zueinander, also durch die Konstruktion der Richtungen (siehe Ende Kapitel 2.5), festgelegt. Richtungen werden als Mittelpunkte von Quadraten auf der Oberfläche des Würfels $[-1, 1]^3$ konstruiert. Aufgrund der unterschiedlichen Positionen der Quadrate auf der Oberfläche des Würfels entstehen durch die Projektion der Richtungen auf die Einheitssphäre unterschiedliche Öffnungswinkel der jeweiligen Kegel. Für jede Stufe $\ell \in \underline{p_{\mathcal{I}}}$ existiert jedoch ein maximaler Öffnungswinkel $2\alpha_\ell \in (0, 2\arctan(\sqrt{2})]$[ii]. Nach der Wahl einer Richtung c wird der Öffnungswinkel als der maximal auftretende Öffnungswinkel zur Stufe der Richtung c gewählt.
Der verwendete Öffnungswinkel ist damit nicht nur unabhängig vom Cluster t, sondern auch unabhängig von dem Cluster s. Dies ermöglicht es, Kegel um eine Richtung $c \notin \mathcal{R}_s$ zu betrachten. Auch wenn dies zunächst nach einer unnützen Eigenschaft klingt, erweist es sich als äußerst praktisch. Denn auf diese Weise kann für den Cluster $s \in \mathcal{T}_{\mathcal{I}}$ mit dem Kind $s' \in \operatorname{kind}(s)$ der Kegel entlang einer Kinderrichtung $c' \in \mathcal{R}_{s'}$ betrachtet werden. Nur Cluster auf der Elternstufe, welche entlang dieser Richtung c' liegen, können auch Kinder hervorbringen, die ebenfalls diese Richtung verwenden würden, denn für den Kegel des Kindes ändert sich nur die Lage der Spitze des Kegels, nicht die Ausbreitungsrichtung des Vektors c'.
Alle Punkte $x \in \mathbb{R}^3$, die im so konstruierten Kegel entlang der Richtung $c \in \mathcal{R}_\ell$ enthalten sind, lassen sich durch die folgenden zwei Bedingungen charakterisieren

$$\frac{\langle p-x, c\rangle_2}{||p-x||_2} \leq \cos(\alpha_\ell),$$
$$\langle p - x, c\rangle_2 \leq ||h_s + h_d||_2,$$

dabei kennzeichnet p die Spitze des Kegels, α_ℓ ist der halbe Öffnungswinkel und $h_s + h_d$ ist die Höhe des Kegels.

[ii] Das Maximum für α_ℓ mit $\arctan(\sqrt{2})$ ergibt sich für den Fall, das pro Würfelseite nur eine Richtung konstruiert wird.

Dass ein so konstruierter Elternkegel auch wirklich alle möglichen Clusterpartner eines seiner Kinder umfasst, ist relativ leicht geometrisch einzusehen. Dazu bezeichne den Elterncluster mit s und ein beliebiges, aber festes seiner Kinder mit $s_i \in \mathrm{kind}(s)$. Die Mittelpunkte der beiden Cluster seien jeweils durch m_s und m_{s_i} gegeben, die betrachteten Kegel werden jeweils mit K_s und K_{s_i} bezeichnet. Beide Kegel werden entlang einer Richtung $c \in \mathcal{R}_{s_i}$ gebildet und durch denselben Öffnungswinkel 2α mit $\alpha = \alpha_{\mathrm{stufe}(s_i)}$ charakterisiert. Der Mittelpunkt m_s des Cluster s liegt auf der Höhe des Kegels K_s, der Mittelpunkt m_{s_i} des Kinderclusters s_i bildet die Spitze des Kegels K_{s_i}. Für K_{s_i} braucht keine Parallelverschiebung der Hypotenuse des zugrundeliegenden Dreiecks durchgeführt werden, da es bei K_{s_i} nicht darum geht, die überdeckenden Quader vollständig zu umfassen, sondern nur darum, die Mittelpunkte der potentiellen Clusterpartner zu beinhalten. Dies ist wichtig für die Argumentation, da der Kinderkegel auf diese Weise ein kleineres Volumen hat und es nur so bei unterschiedlichen Kegelspitzen möglich ist, dass der Elternkegel den Kinderkegel vollständig enthält.

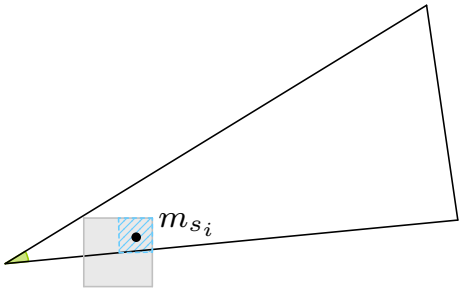


Abbildung 3.5: Der Cluster s_i im Dreieck des Elternkegels

Da der Kegel mit Hilfe eines rechtwinkligen Dreiecks konstruiert wird, entspricht die Höhe des Kegels K_s der Höhe der Dreiecksseite r_k (Vergleich siehe Abbildung 3.4 (b)) und ist damit durch h_s+h_d gegeben. Die Höhe des Kegels K_{s_i} liegt in $(0, h_s+h_d]$, denn der Cluster s_i zieht nur Cluster als potentielle Clusterpartner in Betracht, die beim Elterncluster noch zu den unzulässigen gehörten, also solche, die in der Kugel mit Radius h_d um m_s enthalten sind. Entsprechend liegen die Grundflächen der beiden Kegel K_s und K_{s_i} aufeinander. Der Punkt m_{s_i} liegt im Kegel K_s, denn die Kugel um m_s mit Radius $\frac{1}{2}\operatorname{diam}(Q_s)$ ist vollständig im Kegel K_s enthalten, entsprechend ist m_{s_i} sogar ein innerer Punkt des Kegels K_s.
Der Cluster s_i ist in Abbildung 3.5 blau schraffiert hervorgehoben, der Elterncluster s ist grau hinterlegt, bei dem Dreieck handelt es sich um das dem Kegel K_s zugrundeliegende Dreieck. Die Abbildung 3.6 zeigt das dem Kegel K_{s_i} zugrundeliegende Dreieck in blau

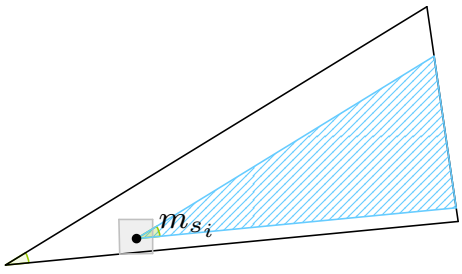


Abbildung 3.6: Den Kegeln zugrundeliegende Dreiecke

schraffiert. Das Bild legt nahe, das K_{s_i} vollständig in K_s enthalten ist, dies kann jedoch auch leicht mit Hilfe von Parallelität und Stufenwinkeln eingesehen werden.

Lemma 3.7
Der Kegel K_{s_i} ist vollständig im Kegel K_s enthalten.

Beweis: Nach den Voraussetzungen liegt m_{s_i} in K_s und beide Kegel haben denselben Öffnungswinkel 2α. Da die Höhen der Kegel entlang der Richtung c gebildet werden, sind die beiden Höhen parallel zueinander. Da m_{s_i} im Inneren von K_s liegt und die Grundflächen der beiden Kegel aufeinander liegen, ist die Höhe des Kegels K_{s_i} vollständig in K_s. Der identische Öffnungswinkel 2α sorgt dafür, dass die Mantelflächen der Kegel ebenfalls parallel zueinander liegen.
Betrachte dazu die zugrundeliegenden Dreiecke der beiden Kegel und verlängere die Hypotenuse des Dreiecks des Kinds s_i über den Punkt m_{s_i} hinaus bis hin zur Höhe $h_s + h_d$. Der entstehende Schnittpunkt ist als q bezeichnet und in Abbildung 3.7 zu finden. Die Höhe $h_s + h_d$ und die verlängerte Seite schneiden sich in q ebenfalls unter dem Winkel α, es handelt sich um Stufenwinkel. Das Stufenwinkelargument liefert beim Betrachten der Hypotenuse des Dreiecks zu s, der verlängerten Hypotenuse des Dreiecks zu s_i und der Höhe $h_s + h_d$, dass die beiden Hypotenusen der zugrundeliegenden Dreiecke parallel zu einander liegen und damit auch die Mantelflächen der Kegel.
Da die Mantelflächen parallel zueinander verlaufen und der Punkt m_{s_i} im Inneren des Kegels K_s liegt, gibt es keinen Punkt aus K_{s_i} der nicht selbst in K_s liegt, denn für solch einen Punkt müssten die Mantelflächen sich schneiden, was durch die Parallelität nicht möglich ist. □

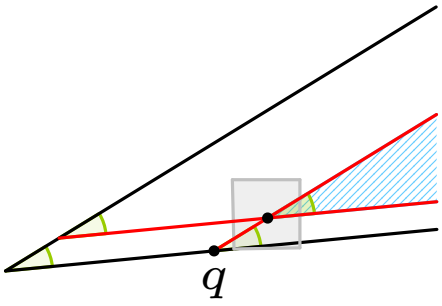


Abbildung 3.7: Verlängerte Seiten und Stufenwinkel

Im letzten Schritt der Dreieckskonstruktion für den Kegel des Elternclusters wurde von dem Mittelpunkt m_s aus auch in Richtung $-c$ um h_s erweitert. Da der Kegel zwar so groß wie nötig, aber so klein wie möglich sein sollte, um eine gute Abschätzung zu erhalten, ist die logische Folgerung hieraus die Verwendung eines Kegelstumpfes $K_{h_k}(s)$ auf der Stufe der Eltern, dem ein Teil der Erweiterung in Richtung $-c$ fehlt. Leider kann nicht die

komplette Erweiterung h_s weggelassen werden, was schnell anhand der Abbildung 3.8 (b), welche eine möglichst ungünstige Kombination zweier Clusterpartner t, s zeigt, einzusehen ist. Aufgrund von unterschiedlichen Durchmessern, der zu untersuchenden überdeckenden Quader sowie einer möglichen Überlappung dieser, kann es vorkommen, dass Teile eines unzulässigen Clusterpartners t entlang der Richtung $-c$ über den Mittelpunkt m_s hinausragen. Dies ist in Abbildung 3.8 (b) durch die gestrichelte Linie angedeutet. Es reicht jedoch aus, entlang der Richtung $-c$ um $\frac{1}{2}\operatorname{diam}_{\max}(\ell)$ zu erweitern, ansonsten würde folgen, dass m_t schon in Richtung $-c$ gelegen haben muss. Entsprechend ergibt sich die Höhe des Kegelstumpfs mit $h_k = h_d + \frac{1}{2}\operatorname{diam}_{\max}(\ell)$.

(a) Kegelstumpf (b) Ungünstige Clusterung

Abbildung 3.8: Kegelstumpf und ungünstige Clusterung

Nachdem geklärt ist, dass ein so konstruierter Kegelstumpf die gewünschte Eigenschaft aufweist, also alle möglichen unzulässigen Clusterpartner entlang einer gegebenen Richtung c enthält, kann es daran gehen, die letzten nötigen Annahmen für die Abschätzung der Anzahl der unzulässigen Clusterpartner zu definieren. Der Kegelstumpf $K_{h_k}(t)$ zu einem Cluster $t \in \mathcal{T}_{\mathcal{I}}^{\ell}$ entlang einer beliebigen Richtung c hat die Höhe h_k, welche mit

$$r_{zb,k}(\ell) := r_{zb}(\ell) + \tfrac{1}{2} \tag{3.1.5}$$

durch

$$h_k = \operatorname{diam}_{\max}(\ell)\, r_{zb,k}(\ell)$$

gegeben ist.
Soll eine Stufe betrachtet werden, auf der nur die Null als Richtung vorhanden ist, muss mit Kugeln statt Kegelstümpfen gearbeitet werden. Für eine Kugel $B_r(m_t)$ um den Mittelpunkt m_t des Clusters $t \in \mathcal{T}_{\mathcal{I}}^{\ell}$ ergibt sich der nötige Radius r der Kugel durch die gleichen Überlegungen wie die Höhe h_d beim Dreieck, welches die Grundlage für den Kegel bildet. Der Radius setzt sich entsprechend aus dem halben Clusterdurchmesser, der Distanz der beiden Cluster und dem maximalen Durchmesser der Cluster auf der Stufe ℓ zusammen und kann erneut mit $r_{zb}(\ell)$ vom konkreten Clusterpaar (t, s) unabhängig gemacht werden. Für den Radius r der Kugel $B_r(m_t)$ ergibt sich

$$r = \operatorname{diam}_{\max}(\ell)\, r_{zb}(\ell).$$

Wie bereits erwähnt, wäre die Verwendung von Kegelstumpf- und Kugelvolumen für die Abschätzungen zu großzügig, da die Cluster sich bei Randelementmethoden allein auf der Oberfläche der Geometrie befinden. Stattdessen werden die Kugel beziehungsweise der Kegelstumpf mit der Oberfläche geschnitten. Dazu nehme an, dass diese Schnitte im Wesentlichen zweidimensional sind, also dass eine Konstante $\mathcal{C}_{Ok} \in \mathbb{R}_{>0}$ existiert, so dass Schnitte der Oberfläche mit einer Kugel durch

$$|B_r(x) \cap \Gamma| \leq \mathcal{C}_{Ok} r^2 \qquad \text{für alle } x \in \mathbb{R}^3, r \in \mathbb{R}_{>0} \tag{3.1.6}$$

beziehungsweise mit einem Kegelstumpf, wobei über die Konstruktion der Richtungen gefolgert werden kann, dass ein $\mathcal{C}_{\overline{D}} \in \mathbb{R}_{>0}$ existiert mit $\#\mathcal{R}_\ell \geq \mathcal{C}_{\overline{D}} \kappa^2 \operatorname{diam}_{\max}(\ell)^2$, durch

$$|K_{h_k}(t) \cap \Gamma| \leq \frac{\mathcal{C}_{Ok}}{\#\mathcal{R}_\ell} h_k^2 \leq \frac{\mathcal{C}_{Ok}}{\mathcal{C}_{\overline{D}} \kappa^2 \operatorname{diam}_{\max}(\ell)^2} h_k^2 \qquad \text{für alle } t \in \mathcal{T}_\mathcal{I}, \ell \in \underline{p_{\mathcal{I}\rfloor}}\, 0, h_k \in \mathbb{R}_{>0} \tag{3.1.7}$$

beschränkt sind. Weiterhin muss die Überlappung von Kugeln um Cluster bedacht werden, also dass gewisse Anteile des Volumens zu mehreren Clustern gehören können. Nehme an, dass sich die Überschneidung von Kugeln um Clustermittelpunkte mit einem Radius, der dem halben Clusterdurchmesser entspricht, durch eine Konstante $\mathcal{C}_{uk} \in \mathbb{N}$

$$\max \# \left\{ t \in \mathcal{T}_\mathcal{I}^\ell \mid x \in B_{\operatorname{diam}(Q_t)/2}(m_t) \right\} \leq \mathcal{C}_{uk} \qquad \text{für alle } x \in \Gamma, \ell \in \underline{p_{\mathcal{I}\rfloor}}\, 0 \tag{3.1.8}$$

abschätzen lässt. Da eine solche Kugel um den Clustermittelpunkt den überdeckenden Quader vollständig enthält, ist $\mathcal{C}_{uk}$ auch eine Schranke für die maximale Anzahl sich überschneidender Quader auf einer Stufe.
Außerdem ist es notwendig, anzunehmen, dass Schnitte einzelner Cluster mit der Oberfläche nicht beliebig klein werden können, also dass eine Konstante $\mathcal{C}_{Ck} \in \mathbb{R}_{>0}$ existiert, so dass

$$|Q_t \cap \Gamma| \geq \frac{\operatorname{diam}^2(Q_t)}{\mathcal{C}_{Ck}} \geq \frac{\operatorname{diam}_{\min}(\ell)^2}{\mathcal{C}_{Ck}} \qquad \text{für alle } \ell \in \underline{p_{\mathcal{I}\rfloor}}\, 0, t \in \mathcal{T}_\mathcal{I}^\ell \tag{3.1.9}$$

gilt. Beliebig kleine Schnitte von Clustern mit der Oberfläche würden bedeuten, dass entweder winzige Cluster konstruiert werden, was die Performance der $\mathcal{RH}^2$-Matrizen gravierend verschlechtern würde, oder dass die Cluster zu sehr großen Teilen gar nicht auf der Oberfläche liegen. Beide Fällen sollten in der Praxis vermieden werden.
Für einen Zeilencluster $t^+ \in \mathcal{T}_\mathcal{I}^{\ell-1}$ und eine Richtung $c \in \mathcal{R}_\ell \setminus \{0\}$ bezeichne mit $U_{t^+}^c$ die Menge der unzulässigen Clusterpartner $s \in \operatorname{row}(t^+)$, die $Q_s \subset K_{h_k}(t^+)$ erfüllen. Entsprechend sei für den richtungsunabhängigen Fall U_{t^+} die Menge der unzulässigen Clusterpartner $s \in \operatorname{row}(t^+)$. Analog definiere $U_{s^+}^c$ beziehungsweise U_{s^+} für einen Spaltencluster $s^+ \in \mathcal{T}_\mathcal{I}^{\ell-1}$ als Teilmenge von $\operatorname{col}(s^+)$. Die Anzahl der unzulässigen Blöcke wird in dem folgenden Lemma abgeschätzt.

Lemma 3.8 (Anzahl unzulässige Blöcke)
Sei ein richtungsabhängiger Blockbaum $\mathcal{T}_{\mathcal{I}\times\mathcal{I}}$ zur Indexmenge $\mathcal{I}$ gegeben. Die Anzahl der unzulässigen Blöcke, die mit einem Cluster $t \in \mathcal{T}_{\mathcal{I}}^{\ell-1}$ für $\ell \in \underline{p_{\mathcal{I}}}$ gebildet werden, kann unter den Annahmen, dass (3.1.9)*,* (3.1.8)*,* (3.1.6) *beziehungsweise* (3.1.7) *erfüllt sind, durch*

$$\#U_t \leq \left(\frac{\operatorname{diam}_{\max}(\ell-1)}{\operatorname{diam}_{\min}(\ell-1)}\right)^2 \mathcal{C}_{uk}\mathcal{C}_{Ck}\mathcal{C}_{Ok} r_{zb}^2(\ell-1)$$

beziehungsweise entlang einer Richtung $c \in \mathcal{R}_\ell \setminus \{0\}$ durch

$$\#U_t^c \leq \left(\frac{\operatorname{diam}_{\max}(\ell-1)}{\operatorname{diam}_{\min}(\ell-1)}\right)^2 \frac{\mathcal{C}_{uk}\mathcal{C}_{Ck}\mathcal{C}_{Ok}}{\#\mathcal{R}_\ell} r_{zb,k}^2(\ell-1) \leq \left(\frac{\operatorname{diam}_{\max}(\ell-1)}{\operatorname{diam}_{\min}(\ell-1)}\right)^2 \frac{\mathcal{C}_{uk}\mathcal{C}_{Ck}\mathcal{C}_{Ok} r_{zb,k}^2(\ell-1)}{\mathcal{C}_{\overline{D}}\kappa^2 \operatorname{diam}_{\max}(\ell)^2}$$

beschränkt werden.

Beweis: Seien ein $\ell \in \underline{p_{\mathcal{I}}}$ und ein Cluster $t \in \mathcal{T}_{\mathcal{I}}^{\ell-1}$ gegeben. Für die Kugel um den Mittelpunkt m_t des Clusters t, die alle unzulässigen Clusterpartner von t umschließt, ist nach obigen Überlegungen ein Radius von $r = \operatorname{diam}_{\max}(\ell-1)\, r_{zb}(\ell-1)$ nötig. Sei weiter Q der überdeckende Quader auf Stufe $\ell-1$, der den kleinsten Schnitt mit der Oberfläche hat. Die maximale Anzahl der unzulässigen Blöcke $\#U_t$ ergibt sich als Quotient aus dem Schnitt der Kugel, die alle unzulässigen Clusterpartner von t umschließt, mit der Oberfläche und dem Schnitt des Quaders Q mit der Oberfläche. Unter Beachtung der Annahmen (3.1.8) zu der Überschneidung von Clustern ergibt sich

$$\begin{aligned}
\#U_t &\leq \frac{\mathcal{C}_{uk}|B_r(m_t)\cap\Gamma|}{|Q\cap\Gamma|} \overset{(3.1.6)}{\leq} \frac{\mathcal{C}_{uk}\mathcal{C}_{Ok}\operatorname{diam}_{\max}(\ell-1)^2 r_{zb}^2(\ell-1)}{|Q\cap\Gamma|} \\
&\overset{(3.1.9)}{\leq} \frac{\mathcal{C}_{uk}\mathcal{C}_{Ck}\mathcal{C}_{Ok}\operatorname{diam}_{\max}(\ell-1)^2 r_{zb}^2(\ell-1)}{\operatorname{diam}_{\min}(\ell-1)^2} \\
&= \left(\frac{\operatorname{diam}_{\max}(\ell-1)}{\operatorname{diam}_{\min}(\ell-1)}\right)^2 \mathcal{C}_{uk}\mathcal{C}_{Ck}\mathcal{C}_{Ok} r_{zb}^2(\ell-1).
\end{aligned}$$

Beim Betrachten eines Kegelstumpfes mit Höhe $h_k = \operatorname{diam}_{\max}(\ell-1)\, r_{zb,k}(\ell-1)$ entlang einer Richtung $c \in \mathcal{R}_\ell$ ergibt sich

$$\#U_t^c \leq \frac{\mathcal{C}_{uk}|K_{h_k}(t)\cap\Gamma|}{|Q\cap\Gamma|}.$$

Dies kann mit der Annahme (3.1.7) weiter abgeschätzt werden

$$\begin{aligned}
\frac{\mathcal{C}_{uk}|K_{h_k}(t)\cap\Gamma|}{|Q\cap\Gamma|} &\leq \frac{\mathcal{C}_{uk}\mathcal{C}_{Ok}\operatorname{diam}_{\max}(\ell-1)^2 r_{zb,k}^2(\ell-1)}{\#\mathcal{R}_\ell|Q\cap\Gamma|} \\
&\overset{(3.1.9)}{\leq} \frac{\mathcal{C}_{uk}\mathcal{C}_{Ck}\mathcal{C}_{Ok}\operatorname{diam}_{\max}(\ell-1)^2 r_{zb,k}^2(\ell-1)}{\#\mathcal{R}_\ell \operatorname{diam}_{\min}(\ell-1)^2},
\end{aligned}$$

schließlich kann $\#\mathcal{R}_\ell$ noch nach unten durch $\mathcal{C}_{\overline{D}}\kappa^2 \operatorname{diam}_{\max}(\ell)^2$ beschränkt werden. □

Da beim Bilden der Clusterbäume ein Cluster $t \in \mathcal{T}_{\mathcal{I}}$ theoretisch viele Kinder haben kann, ist es sinnvoll, die maximale Anzahl an auftretenden Kindern festzuhalten. Diese Zahl wird auch benötigt, um aus der Anzahl der unzulässigen Clusterpartner $\#U_t$ beziehungsweise $\#U_t^c$ eine Schranke für die möglichen Clusterpartner der Kinderstufe zu gewinnen.

Definition 3.9 (Kinderkonstante)
Definiere zu einem gegebenen richtungsabhängigen Clusterbaum $\mathcal{T}_\mathcal{I}$ die Kinderkonstante $\mathcal{C}_{kk} \in \mathbb{N}_{>1}$ *durch*

$$\mathcal{C}_{kk} := \max \{ \# \operatorname{kind}(t) \mid t \in \mathcal{T}_\mathcal{I} \}.$$

Damit ist alles zusammen, um eine Aussage zur stufenweisen Schwachbesetztheit nachzuweisen.

Satz 3.10 (Schwachbesetztheit)
Sei ein richtungsabhängiger Blockbaum $\mathcal{T}_{\mathcal{I}\times\mathcal{I}}$ zur Indexmenge $\mathcal{I}$ gegeben. Unter der Voraussetzung von Lemma 3.8 gilt für jedes $t \in \mathcal{T}_\mathcal{I} \setminus \{\operatorname{wurzel}(\mathcal{T}_\mathcal{I})\}$ mit $\operatorname{stufe}(t) = \ell$

$$\# \operatorname{row}(t), \# \operatorname{col}(t) \leq \mathcal{C}_{kk}\mathcal{C}_{uk}\mathcal{C}_{Ck}\mathcal{C}_{Ok} \left(\frac{\operatorname{diam}_{\max}(\ell-1)}{\operatorname{diam}_{\min}(\ell-1)} \right)^2 r_{zb}^2(\ell-1),$$

insbesondere gilt für $c \in \mathcal{R}_\ell \setminus \{0\}$

$$\# \operatorname{row}_c(t), \# \operatorname{col}_c(t) \leq \mathcal{C}_{kk}\mathcal{C}_{uk}\mathcal{C}_{Ck}\mathcal{C}_{Ok} \left(\frac{\operatorname{diam}_{\max}(\ell-1)}{\operatorname{diam}_{\min}(\ell-1)} \right)^2 \frac{r_{zb,k}^2(\ell-1)}{\mathcal{C}_{\overline{D}}\kappa^2 \operatorname{diam}_{\max}(\ell)^2}.$$

Beweis: Betrachte allein $\operatorname{row}(t)$, die Aussage folgt aufgrund der Symmetrie der Zulässigkeitsbedingungen analog für $\operatorname{col}(t)$.
Sei t^+ der Elterncluster von t. Mit den unzulässigen Clusterpartnern des Elternclusters gilt

$$\operatorname{row}(t) \subset \bigcup_{s^+ \in U_{t^+}} \operatorname{kind}(s^+).$$

Mit der Kinderkonstante (siehe Definition 3.9) kann die Mächtigkeit von $\operatorname{row}(t)$ wie folgt abgeschätzt werden

$$\# \operatorname{row}(t) \leq \sum_{s^+ \in U_{t^+}} \# \operatorname{kind}(s^+) \leq \mathcal{C}_{kk} \sum_{s^+ \in U_{t^+}} 1 \leq \mathcal{C}_{kk} \# U_{t^+}.$$

Es folgt mit Lemma 3.8

$$\# \operatorname{row}(t) \leq \mathcal{C}_{kk}\mathcal{C}_{uk}\mathcal{C}_{Ck}\mathcal{C}_{Ok} \left(\frac{\operatorname{diam}_{\max}(\ell-1)}{\operatorname{diam}_{\min}(\ell-1)} \right)^2 r_{zb}^2(\ell-1).$$

Angenommen es gilt $\#\mathcal{R}_\ell > 1$. Dann kann auch eine Abschätzung für $\operatorname{row}_c(t)$ gewonnen werden. In diesem Fall müssen nur Clusterpartner auf der Stufe davor betrachtet werden, die innerhalb des Kegels um die betrachtete Richtung c liegen, da nur deren Kinder auf Stufe ℓ die Richtung c verwenden können, wie anhand der Kegel gezeigt wurde.
Sei eine Richtung $c \in \mathcal{R}_\ell \setminus \{0\}$ gegeben, dann folgt für $\operatorname{row}_c(t)$

$$\operatorname{row}_c(t) \subset \bigcup_{s^+ \in U_{t^+}^c} \operatorname{kind}(s^+).$$

Mit der Kinderkonstante und dem Lemma 3.8 kann die Mächtigkeit von $\mathrm{row}_c(t)$ mit

$$\begin{aligned}\#\,\mathrm{row}_c(t) &\leq \sum_{s^+\in U^c_{t^+}} \#\,\mathrm{kind}(s^+) = \mathcal{C}_{kk} \sum_{s^+\in U^c_{t^+}} 1 \leq \mathcal{C}_{kk}\#U^c_{t^+} \\ &\leq \mathcal{C}_{kk}\mathcal{C}_{uk}\mathcal{C}_{Ck}\mathcal{C}_{Ok}\left(\frac{\mathrm{diam}_{\max}(\ell-1)}{\mathrm{diam}_{\min}(\ell-1)}\right)^2 \frac{r^2_{zb,k}(\ell-1)}{\mathcal{C}_{\overline{D}}\kappa^2\,\mathrm{diam}_{\max}(\ell)^2}\end{aligned}$$

abgeschätzt werden. □

Im letzten Satz ist die Wurzel des Baums von der Betrachtung ausgeschlossen gewesen, für sie gilt jedoch trivialerweise $\#\,\mathrm{row}(t) = 1$.

Korollar 3.11
Die Aussage des Satzes 3.10 kann für einen Clusters $t \in \mathcal{T}_\mathcal{I} \setminus \{\mathrm{wurzel}\,\mathcal{T}_\mathcal{I}\}$ mit $\mathrm{stufe}(t) = \ell$, falls $\#\mathcal{R}_\ell > 1$ gilt, auch durch

$$\#\,\mathrm{row}(t), \#\,\mathrm{col}(t) \leq \mathcal{C}_{kk}\mathcal{C}_{uk}\mathcal{C}_{Ck}\mathcal{C}_{Ok}\frac{\#\mathcal{R}^{eff}_t}{\#\mathcal{R}_\ell}\left(\frac{\mathrm{diam}_{\max}(\ell-1)}{\mathrm{diam}_{\min}(\ell-1)}\right)^2 r^2_{zb,k}(\ell-1)$$

angegeben werden.

Beweis: Mit Satz 3.10, wobei $\#\mathcal{R}_\ell$ noch nicht nach unten abgeschätzt wurde, folgt für die Mächtigkeit von $\mathrm{row}(t)$ auch

$$\begin{aligned}\#\,\mathrm{row}(t) &= \sum_{c\in\mathcal{R}^{eff}_t} \#\,\mathrm{row}_c\,(t) \leq \#\mathcal{R}^{eff}_t\mathcal{C}_{kk}\#U^c_{t^+} \\ &\leq \mathcal{C}_{kk}\mathcal{C}_{uk}\mathcal{C}_{Ck}\mathcal{C}_{Ok}\frac{\#\mathcal{R}^{eff}_t}{\#\mathcal{R}_\ell}\left(\frac{\mathrm{diam}_{\max}(\ell-1)}{\mathrm{diam}_{\min}(\ell-1)}\right)^2 r^2_{zb,k}(\ell-1).\end{aligned}$$

□

Für $\mathrm{stufe}(t) \geq \ell$ gilt immer $\frac{\#\mathcal{R}^{eff}_t}{\#\mathcal{R}_\ell} \leq 1$, wenn zusätzlich noch

$$\frac{\#\mathcal{R}^{eff}_t}{\#\mathcal{R}_\ell} r^2_{zb,k}(\ell-1) \leq r^2_{zb}(\ell-1)$$

erfüllt ist, führt die Abschätzung mit Hilfe der Kegelstümpfe zu durchaus schärferen Aussagen als mit Kugeln allein. Der Ansatz mit Kugeln allein liefert jedoch grundsätzlich eine obere Schranke für die Anzahl der möglichen Clusterpartner für den betrachteten Cluster, weshalb im Folgenden allein mit dieser Variante weiter gearbeitet wird.
Für weitere Abschätzungen erweist sich das Verhältnis von maximalen zu minimalen Clusterdurchmessern als unhandlich. Durch die Wahl, wie die überdeckenden Quader erstellt werden, ist $\frac{\mathrm{diam}_{\max}(\ell-1)}{\mathrm{diam}_{\min}(\ell-1)}$ jedoch kontrollierbar (bei gleichgroßen Quadern wird dies zu 1). Entsprechend kann davon ausgegangen werden, dass eine Konstante $\mathcal{C}_{mk} \in \mathbb{R}_{\geq 1}$ existiert, so dass

$$\frac{\mathrm{diam}_{\max}(\ell)}{\mathrm{diam}_{\min}(\ell)} \leq \mathcal{C}_{mk} \qquad \text{für alle } \ell \in \underline{p_{\mathcal{I}}} \tag{3.1.10}$$

gilt. Dies ermöglicht es, eine ähnliche Aussage zur Schwachbesetztheit wie in [6, Lem. 8] zu machen.

Korollar 3.12
Die Schwachbesetztheit nach Satz 3.10 kann unter der zusätzlichen Annahme von (3.1.10) *für alle Cluster $t \in \mathcal{T}_{\mathcal{I}}^{\ell}$ mit $\ell \in \underline{p_{\mathcal{I}|}}$ durch*

$$\#\operatorname{row}(t), \#\operatorname{col}(t) \leq \begin{cases} \mathcal{C}_{sk} & \text{falls } \kappa \operatorname{diam}_{\max}(\ell-1) \leq 1, \\ \mathcal{C}_{sk}\kappa^2 \operatorname{diam}_{\max}(\ell-1)^2 & sonst, \end{cases}$$

wobei

$$\mathcal{C}_{sk} = \mathcal{C}_{kk}\mathcal{C}_{uk}\mathcal{C}_{Ck}\mathcal{C}_{Ok}\mathcal{C}_{mk}^2 \left(\tfrac{3}{2} + \tfrac{1}{\eta_2}\right)^2 \in \mathbb{R}_{\geq 1}$$

erfüllt, weiter abgeschätzt werden.

Beweis: Im Fall, dass $\kappa \operatorname{diam}_{\max}(\ell-1) > 1$ gilt, kann $r_{zb}(\ell-1)$ wegen

$$\begin{aligned} \tfrac{3}{2} + \tfrac{\operatorname{diam}_{\max}(\ell-1)\,\kappa}{\eta_2} &\leq \tfrac{3\operatorname{diam}_{\max}(\ell-1)\,\kappa}{2} + \tfrac{\operatorname{diam}_{\max}(\ell-1)\,\kappa}{\eta_2} \\ &= \operatorname{diam}_{\max}(\ell-1)\,\kappa \left(\tfrac{3}{2} + \tfrac{1}{\eta_2}\right) \end{aligned}$$

auch mit

$$r_{zb}(\ell-1) \leq \operatorname{diam}_{\max}(\ell-1)\,\kappa \left(\tfrac{3}{2} + \tfrac{1}{\eta_2}\right)$$

beschränkt werden. Falls $\kappa \operatorname{diam}_{\max}(\ell-1) \leq 1$ erfüllt ist, folgt schon aus der Definition von $r_{zb}(\ell-1)$, dass

$$r_{zb}(\ell-1) = \tfrac{3}{2} + \tfrac{1}{\eta_2}$$

gilt. Mit der Annahme (3.1.10) kann auch der Quotient der Durchmesser durch

$$\left(\tfrac{\operatorname{diam}_{\max}(\ell-1)}{\operatorname{diam}_{\min}(\ell-1)}\right)^2 \leq \mathcal{C}_{mk}^2$$

beschränkt werden, womit sich die Behauptung ergibt. □

Zunächst kann noch eine Aussage zur Anzahl der Cluster auf einer Stufe im Baum gemacht werden.

Lemma 3.13 (Anzahl Cluster auf einer Stufe)
Sei ein richtungsabhängiger Clusterbaum $\mathcal{T}_{\mathcal{I}}$ zur Indexmenge $\mathcal{I}$ gegeben. Unter den Voraussetzungen von (3.1.9) *und* (3.1.8) *gilt für alle $\ell \in \underline{p_{\mathcal{I}|}}$*

$$\#\mathcal{T}_{\mathcal{I}}^{\ell} \leq \tfrac{\mathcal{C}_{Ck}\mathcal{C}_{uk}|\Gamma|}{(\operatorname{diam}_{\min}(\ell))^2}. \tag{3.1.11}$$

Beweis: Für alle $\ell \in \underline{p_{\mathcal{I}|}}$ gilt

$$\begin{aligned}(\operatorname{diam}_{\min}(\ell))^2 \#\mathcal{T}_{\mathcal{I}}^{\ell} &\leq \sum_{t\in\mathcal{T}_{\mathcal{I}}^{\ell}} \operatorname{diam}^2(Q_t) \overset{(3.1.9)}{\leq} \sum_{t\in\mathcal{T}_{\mathcal{I}}^{\ell}} \mathcal{C}_{Ck}|Q_t\cap\Gamma| \\ &= \mathcal{C}_{Ck}\int_{\Gamma}\sum_{t\in\mathcal{T}_{\mathcal{I}}^{\ell}} 1_{Q_t}(x)\,\mathrm{d}x \leq \mathcal{C}_{Ck}\int_{\Gamma}\sum_{t\in\mathcal{T}_{\mathcal{I}}^{\ell}} 1_{B_{\operatorname{diam}(Q_t)/2}(m_t)}(x)\,\mathrm{d}x \\ &\leq \mathcal{C}_{Ck}\mathcal{C}_{uk}\int_{\Gamma} 1\,\mathrm{d}x = \mathcal{C}_{Ck}\mathcal{C}_{uk}|\Gamma|.\end{aligned}$$

□[6]

Diese Aussage kann auch genutzt werden, um die Anzahl der Cluster für alle auftretenden Richtungen abzuschätzen.

Lemma 3.14 (Anzahl der richtungsabhängigen Cluster)
Sei ein richtungsabhängiger Blockbaum $\mathcal{T}_{\mathcal{I}\times\mathcal{I}}$ zur Indexmenge $\mathcal{I}$ gegeben. Unter den Voraussetzungen von Lemma 3.13 sowie (3.1.10) *und* (3.1.4) *gilt*

$$\sum_{t\in\mathcal{T}_{\mathcal{I}}}\sum_{c\in\mathcal{R}_t^{eff}} 1 \leq \#\mathcal{T}_{\mathcal{I}} + \kappa^2(p_{\mathcal{I}}+1)\mathcal{C}_{tk}, \tag{3.1.12}$$

wobei sich $\mathcal{C}_{tk}$ wie folgt zusammensetzt

$$\mathcal{C}_{tk} := \mathcal{C}_{\mathcal{D}}\mathcal{C}_{Ck}\mathcal{C}_{uk}|\Gamma|\mathcal{C}_{mk}^2.$$

Beweis: Es gilt

$$\begin{aligned}\sum_{t\in\mathcal{T}_{\mathcal{I}}}\sum_{c\in\mathcal{R}_t^{eff}} 1 &= \sum_{\ell=0}^{p_{\mathcal{I}}}\sum_{t\in\mathcal{T}_{\mathcal{I}}^{\ell}} \#\mathcal{R}_t^{eff} \\ &\overset{(3.1.4)}{\leq} \sum_{\ell=0}^{p_{\mathcal{I}}}\sum_{t\in\mathcal{T}_{\mathcal{I}}^{\ell}} (1+\kappa^2\mathcal{C}_{\mathcal{D}}\operatorname{diam}_{\max}(\ell)^2) \\ &= \#\mathcal{T}_{\mathcal{I}} + \sum_{\ell=0}^{p_{\mathcal{I}}} \kappa^2\mathcal{C}_{\mathcal{D}}\operatorname{diam}_{\max}(\ell)^2\,\#\mathcal{T}_{\mathcal{I}}^{\ell} \\ &\overset{3.13}{\leq} \#\mathcal{T}_{\mathcal{I}} + \sum_{\ell=0}^{p_{\mathcal{I}}} \kappa^2\mathcal{C}_{\mathcal{D}}\operatorname{diam}_{\max}(\ell)^2\,\tfrac{\mathcal{C}_{Ck}\mathcal{C}_{uk}|\Gamma|}{(\operatorname{diam}_{\min}(\ell))^2} \\ &\overset{(3.1.10)}{\leq} \#\mathcal{T}_{\mathcal{I}} + \kappa^2\mathcal{C}_{\mathcal{D}}\mathcal{C}_{Ck}\mathcal{C}_{uk}|\Gamma|\mathcal{C}_{mk}^2\sum_{\ell=0}^{p_{\mathcal{I}}} 1 \\ &= \#\mathcal{T}_{\mathcal{I}} + \kappa^2\mathcal{C}_{\mathcal{D}}\mathcal{C}_{Ck}\mathcal{C}_{uk}|\Gamma|\mathcal{C}_{mk}^2(p_{\mathcal{I}}+1).\end{aligned}$$

□[10]

Mit Lemma 3.13 zusammen mit der Schwachbesetztheit können auch Summen der Form

$$\sum_{t\in\mathcal{T}_\mathcal{I}} \sum_{c\in\mathcal{R}_t^{eff}} \#\operatorname{row}_c(t)$$

beschränkt werden.

Lemma 3.15 (Anzahl Blöcke)
Sei ein richtungsabhängiger Blockbaum $\mathcal{T}_{\mathcal{I}\times\mathcal{I}}$ zur Indexmenge $\mathcal{I}$ gegeben. Unter den Voraussetzungen von Lemma 3.8 und mit (3.1.10) *gilt*

$$\sum_{t\in\mathcal{T}_\mathcal{I}} \sum_{c\in\mathcal{R}_t^{eff}} \#\operatorname{row}_c(t) \leq \mathcal{C}_{sk}\left(\#\mathcal{T}_\mathcal{I} + p_\mathcal{I}\kappa^2\mathcal{C}_{kk}\mathcal{C}_{Ck}\mathcal{C}_{uk}\mathcal{C}_{mk}^2|\Gamma|\right) + 1.$$

Beweis: Es gilt

$$\sum_{t\in\mathcal{T}_\mathcal{I}} \sum_{c\in\mathcal{R}_t^{eff}} \#\operatorname{row}_c(t) \leq \sum_{t\in\mathcal{T}_\mathcal{I}} \#\operatorname{row}(t) = \sum_{\ell=1}^{p_\mathcal{I}} \sum_{t\in\mathcal{T}_\mathcal{I}^\ell} \#\operatorname{row}(t) + 1,$$

denn für die Wurzel des Baums ist $\#\operatorname{row}(t) = 1$ erfüllt. Die Schwachbesetztheit hängt vom Elterncluster t^+ und dessen Durchmesser ab, so dass mit Korollar 3.12

$$\begin{aligned}
&\sum_{\ell=1}^{p_\mathcal{I}} \sum_{t\in\mathcal{T}_\mathcal{I}^\ell} \#\operatorname{row}(t) + 1 \\
&\quad\leq \mathcal{C}_{sk}\left(\sum_{\ell=1}^{p_\mathcal{I}} \sum_{\substack{t\in\mathcal{T}_\mathcal{I}^\ell \\ \kappa\operatorname{diam}_{\max}(\ell-1)\leq 1}} 1 + \sum_{\ell=1}^{p_\mathcal{I}} \sum_{\substack{t\in\mathcal{T}_\mathcal{I}^\ell \\ \kappa\operatorname{diam}_{\max}(\ell-1)>1}} \kappa^2(\operatorname{diam}_{\max}(\ell-1))^2\right) + 1 \\
&\quad\leq \mathcal{C}_{sk}\left(\#\mathcal{T}_\mathcal{I} + \sum_{\ell=0}^{p_\mathcal{I}-1} \sum_{\substack{t^+\in\mathcal{T}_\mathcal{I}^\ell \\ \kappa\operatorname{diam}_{\max}(\ell)>1}} \mathcal{C}_{kk}\kappa^2(\operatorname{diam}_{\max}(\ell))^2\right) + 1
\end{aligned}$$

folgt. Die verbleibende Doppelsumme kann mit (3.1.11) und (3.1.10) abgeschätzt werden

$$\begin{aligned}
\mathcal{C}_{kk}\sum_{\ell=0}^{p_\mathcal{I}-1} \sum_{\substack{t^+\in\mathcal{T}_\mathcal{I}^\ell \\ \kappa\operatorname{diam}_{\max}(\ell)>1}} \kappa^2(\operatorname{diam}_{\max}(\ell))^2 &\leq \mathcal{C}_{kk}\sum_{\ell=0}^{p_\mathcal{I}-1} \kappa^2(\operatorname{diam}_{\max}(\ell))^2\#\mathcal{T}_\mathcal{I}^\ell \\
&\leq \mathcal{C}_{kk}\sum_{\ell=0}^{p_\mathcal{I}-1} \kappa^2(\operatorname{diam}_{\max}(\ell))^2\frac{\mathcal{C}_{Ck}\mathcal{C}_{uk}|\Gamma|}{(\operatorname{diam}_{\min}(\ell))^2} \\
&\leq p_\mathcal{I}\kappa^2\mathcal{C}_{kk}\mathcal{C}_{Ck}\mathcal{C}_{uk}\mathcal{C}_{mk}^2|\Gamma|.
\end{aligned}$$

Insgesamt ergibt sich so eine Abschätzung von

$$\sum_{t\in\mathcal{T}_\mathcal{I}} \sum_{c\in\mathcal{R}_t^{eff}} \#\operatorname{row}_c(t) \leq \mathcal{C}_{sk}\left(\#\mathcal{T}_\mathcal{I} + p_\mathcal{I}\kappa^2\mathcal{C}_{kk}\mathcal{C}_{Ck}\mathcal{C}_{uk}\mathcal{C}_{mk}^2|\Gamma|\right) + 1.$$

□[10]

Bevor es endlich an die erste Komplexitätsabschätzung geht, wird noch eine weitere Konstante eingeführt, um mit ihr die Anzahl aller Blätter und aller Cluster eines Clusterbaums abschätzen zu können. Nehme an, dass die Größe der Blattcluster nicht zu stark variiert, also dass eine Konstante $\mathcal{C}_{bk} \in \mathbb{R}_{>0}$ existiert, die zusammen mit dem maximalen Rang k (siehe Definition 3.5)

$$\mathcal{C}_{bk}^{-1}k \leq \#^\mathfrak{I}t \leq \mathcal{C}_{bk}k \qquad \text{für alle } t \in \mathcal{L}_\mathcal{I} \tag{3.1.13}$$

erfüllt.

Bemerkung 17 (Clusterauflösung): *In der Praxis kann die Auflösung für den Clusteralgorithmus festgelegt werden. Die Wahl der Auflösung sollte dabei an den Rang k gekoppelt sein. Bei der Interpolation haben sich $2|M|^3 = k$ sowie $2|M|^2 = k$ bewährt, wobei $|M| = m + 1$ die Anzahl der Stützstellen in einer Dimension ist. Je nach Geometrie kann es aber immer zu Clustern mit wenigen bis hin zu einem einzigen Element kommen, so dass eine generell ‚optimale' Wahl der Auflösung nicht möglich ist. Der Wert der Konstante $\mathcal{C}_{bk}$ hängt damit von der Wahl der Auflösung, der Geometrie und der Clusterstrategie ab.*

Lemma 3.16 (Anzahl der Blätter)
Sei ein richtungsabhängiger Clusterbaum $\mathcal{T}_\mathcal{I}$ zur Indexmenge $\mathcal{I}$ gegeben. Unter der Voraussetzung von (3.1.13) *kann die Anzahl der Blattcluster beschränkt werden mit*

$$\#\mathcal{L}_\mathcal{I} \leq \tfrac{\mathcal{C}_{bk}}{k}\#\mathcal{I}.$$

Beweis: Mit Lemma 2.22 gilt

$$\#\mathcal{L}_\mathcal{I} = \sum_{t\in\mathcal{L}_\mathcal{I}} 1 \leq \tfrac{\mathcal{C}_{bk}}{k} \sum_{t\in\mathcal{L}_\mathcal{I}} \#^\mathfrak{I}t = \tfrac{\mathcal{C}_{bk}}{k}\#\mathcal{I}.$$

□[6,*Lem.*9]

Korollar 3.17 (Anzahl der Cluster)
Unter der Voraussetzung von Lemma (3.16) *kann die Anzahl der Cluster $\#\mathcal{T}_\mathcal{I}$ beschränkt werden mit*

$$\#\mathcal{T}_\mathcal{I} \leq \tfrac{2\mathcal{C}_{bk}}{k}\#\mathcal{I}.$$

Beweis: Aus Lemma 2.23 und Lemma (3.16) folgt unmittelbar

$$\#\mathcal{T}_{\mathcal{I}} \leq 2\#\mathcal{L}_{\mathcal{I}} - 1 \leq \tfrac{2\mathcal{C}_{bk}}{k}\#\mathcal{I}.$$

$\square$[6,*Lem. 9*]

Weiterhin ist eine Regularitätsbedingung für die Anzahl der Elemente in unzulässigen Blöcken nötig, das heißt, dass die beteiligten Cluster ähnlich viele Indizes enthalten sollen. Falls $t \in \mathcal{L}_{\mathcal{I}}$, $s \in \mathcal{T}_{\mathcal{I}}$ und

$$\begin{aligned} &\eta_2 \operatorname{dist}(Q_t, Q_s) < \kappa \max\left\{\operatorname{diam}^2(Q_t), \operatorname{diam}^2(Q_s)\right\} \qquad \text{oder} \\ &\eta_2 \operatorname{dist}(Q_t, Q_s) < \max\left\{\operatorname{diam}(Q_t), \operatorname{diam}(Q_s)\right\} \end{aligned}$$

erfüllt ist, gelte

$$\#^{\mathfrak{I}} s \leq \mathcal{C}_{gk} \#^{\mathfrak{I}} t \tag{3.1.14}$$

für eine Konstante $\mathcal{C}_{gk} \in \mathbb{R}_{>0}$.

Mit dieser letzten Annahme soll es an die Aufwandsabschätzung gehen. Beginne mit dem Speicherbedarf für unzulässige Blätter und die Kopplungsmatrizen, deren Beweis sich an [6, Lem. 11] orientiert.

Lemma 3.18 (Speicheraufwand Nah-und Fernfeld)
Sei ein richtungsabhängiger Blockbaum $\mathcal{T}_{\mathcal{I}\times\mathcal{I}}$ zur Indexmenge $\mathcal{I}$ gegeben. Unter den Voraussetzungen von Lemma 3.8 sowie (3.1.10)*,* (3.1.13) *und* (3.1.14) *beläuft sich der Aufwand, die Nah- und Fernfeldmatrizen zu speichern, auf*

$$\mathcal{C}_{nuf}(k\#\mathcal{I} + k^2(p_{\mathcal{I}}\kappa^2 + 1)),$$

wobei $\mathcal{C}_{nuf} := \mathcal{C}_{nn} \max\left\{2\mathcal{C}_{sk}\mathcal{C}_{bk}, \mathcal{C}_{sk}\mathcal{C}_{kk}\mathcal{C}_{Ck}\mathcal{C}_{uk}\mathcal{C}_{mk}^2|\Gamma|, 1\right\}$ mit der Konstanten $\mathcal{C}_{nn} := \max\left\{1, \mathcal{C}_{gk}\mathcal{C}_{bk}^2\right\}$ ist.

Beweis: Betrachte zunächst die Nahfeldmatrizen, sei dazu $b = (t, s, c)$ ein unzulässiger Block. Dann müssen $(\#^{\mathfrak{I}} t)(\#^{\mathfrak{I}} s)$ Einträge gespeichert werden. Da es sich um einen unzulässigen Block handelt, muss es sich bei t oder s um ein Blatt handeln. Ohne Beschränkung der Allgemeinheit nehme an, dass t das Blatt ist, dann gilt mit (3.1.14) und (3.1.13)

$$(\#^{\mathfrak{I}} t)(\#^{\mathfrak{I}} s) \leq \mathcal{C}_{gk}(\#^{\mathfrak{I}} t)^2 \leq \mathcal{C}_{gk}\mathcal{C}_{bk}^2 k^2.$$

Entsprechend ergibt sich für die Gesamtheit der Nahfeldmatrizen

$$\sum_{b \in \mathcal{L}^-_{\mathcal{I}\times\mathcal{I}}} (\#^{\mathfrak{I}} t)(\#^{\mathfrak{I}} s) \leq \sum_{b \in \mathcal{L}^-_{\mathcal{I}\times\mathcal{I}}} \mathcal{C}_{gk}\mathcal{C}_{bk}^2 k^2,$$

während für die Kopplungsmatrizen ein Speicher von

$$\sum_{b\in\mathcal{L}^+_{\mathcal{I}\times\mathcal{I}}} k_{tc}k_{sc} \leq \sum_{b\in\mathcal{L}^+_{\mathcal{I}\times\mathcal{I}}} k^2$$

anfällt. Insgesamt führt dies mit Lemma 3.15 und Korollar 3.17 sowie $\mathcal{C}_{nn}$ zu einem Speicheraufwand von

$$\begin{aligned}
\sum_{b\in\mathcal{L}^-_{\mathcal{I}\times\mathcal{I}}} \mathcal{C}_{gk}\mathcal{C}_{bk}^2k^2 + \sum_{b\in\mathcal{L}^+_{\mathcal{I}\times\mathcal{I}}} k^2 &\leq \sum_{b\in\mathcal{L}_{\mathcal{I}\times\mathcal{I}}} \mathcal{C}_{nn}k^2 \leq \mathcal{C}_{nn}k^2 \sum_{t\in\mathcal{T}_\mathcal{I}} \operatorname{row}(t)\\
&\leq \mathcal{C}_{nn}k^2\left(\mathcal{C}_{sk}\left(\#\mathcal{T}_\mathcal{I} + p_\mathcal{I}\kappa^2\mathcal{C}_{kk}\mathcal{C}_{Ck}\mathcal{C}_{uk}\mathcal{C}_{mk}^2|\Gamma|\right)+1\right)\\
&\leq \mathcal{C}_{nn}k^2\left(\mathcal{C}_{sk}(\tfrac{2\mathcal{C}_{bk}}{k}\#\mathcal{I} + p_\mathcal{I}\kappa^2\mathcal{C}_{kk}\mathcal{C}_{Ck}\mathcal{C}_{uk}\mathcal{C}_{mk}^2|\Gamma|)+1\right)\\
&= \mathcal{C}_{nn}\left(2k\mathcal{C}_{sk}\mathcal{C}_{bk}\#\mathcal{I} + k^2(p_\mathcal{I}\kappa^2\mathcal{C}_{sk}\mathcal{C}_{kk}\mathcal{C}_{Ck}\mathcal{C}_{uk}\mathcal{C}_{mk}^2|\Gamma|+1)\right)\\
&\leq \mathcal{C}_{nuf}\left(k\#\mathcal{I} + k^2(p_\mathcal{I}\kappa^2+1)\right).
\end{aligned}$$

□

Um den gesamten Speicherbedarf angeben zu können, fehlen noch die Clusterbasen.

Lemma 3.19 (Speicherbedarf Clusterbasis)
Sei ein richtungsabhängiger Blockbaum $\mathcal{T}_{\mathcal{I}\times\mathcal{I}}$ zur Indexmenge $\mathcal{I}$ gegeben. Unter den Voraussetzungen von Lemma 3.13 sowie (3.1.4),(3.1.10) *und* (3.1.13) *ist der Speicherbedarf für eine richtungsabhängige Clusterbasis zum Clusterbaum $\mathcal{T}_\mathcal{I}$ mit der Familie an Richtungsmengen $\mathcal{R}$ beschränkt durch*

$$k\mathcal{C}_{cb}\left(\#\mathcal{I} + k\kappa^2(p_\mathcal{I}+1)\right),$$

wobei

$$C_{cb} := \max\{\mathcal{C}_{kk},\mathcal{C}_{bk}\}\max\{2\mathcal{C}_{bk},\mathcal{C}_{tk}\}$$

gilt.

Beweis: Zunächst betrachte den Fall $t \in \mathcal{L}_\mathcal{I}$, dann muss für jede verwendete Richtung eine Matrix mit maximal $(\#^\mathfrak{I} t)k_{tc}$ Einträgen gespeichert werden. Dies führt mit (3.1.13) zu

$$\sum_{t\in\mathcal{L}_\mathcal{I}}\sum_{c\in\mathcal{R}_t^{eff}} (\#^\mathfrak{I} t)k_{tc} \leq \sum_{t\in\mathcal{L}_\mathcal{I}}\sum_{c\in\mathcal{R}_t^{eff}} k^2\mathcal{C}_{bk} \leq \sum_{t\in\mathcal{L}_\mathcal{I}} \#\mathcal{R}_t^{eff}k^2\mathcal{C}_{bk}.$$

Falls $t \in \mathcal{T}_\mathcal{I}\setminus\mathcal{L}_\mathcal{I}$ erfüllt, wird für jede benötigte Richtung c eine kleine Transfermatrix zum Kind t' mit Richtung c' gespeichert, was pro Matrix $k_{tc}k_{t'c'}$ Speicherplätze erfordert und insgesamt zu

$$\sum_{t\in\mathcal{T}_\mathcal{I}\setminus\mathcal{L}_\mathcal{I}}\sum_{c\in\mathcal{R}_t^{eff}}\sum_{t'\in\operatorname{kind}(t)} k_{tc}k_{t'c'} \leq \sum_{t\in\mathcal{T}_\mathcal{I}\setminus\mathcal{L}_\mathcal{I}}\sum_{c\in\mathcal{R}_t^{eff}} k^2\mathcal{C}_{kk} \leq \sum_{t\in\mathcal{T}_\mathcal{I}\setminus\mathcal{L}_\mathcal{I}} \#\mathcal{R}_t^{eff}k^2\mathcal{C}_{kk}$$

führt. Mit $\widehat{\mathcal{C}}_{cb} = \max\{\mathcal{C}_{kk}, \mathcal{C}_{bk}\}$ ist der Speicherbedarf für die gesamte Clusterbasis beschränkt durch

$$k^2\widehat{\mathcal{C}}_{cb}\sum_{t\in\mathcal{T}_\mathcal{I}}\#\mathcal{R}_t^{eff} \overset{3.14}{\leq} k^2\widehat{\mathcal{C}}_{cb}\left(\#\mathcal{T}_\mathcal{I} + \kappa^2(p_\mathcal{I}+1)\mathcal{C}_{tk}\right) \overset{3.17}{\leq} k\widehat{\mathcal{C}}_{cb}\left(2\mathcal{C}_{bk}\#\mathcal{I} + k\kappa^2(p_\mathcal{I}+1)\mathcal{C}_{tk}\right).$$

□

Der Aufwand liegt damit in $\mathcal{O}(k\#\mathcal{I} + k^2\kappa^2\log_2(\#\mathcal{I}))$.

Theorem 3.20 (Speicheraufwand $\mathcal{RH}^2$-Matrix)
Sei ein richtungsabhängiger Blockbaum $\mathcal{T}_{\mathcal{I}\times\mathcal{I}}$ zur Indexmenge $\mathcal{I}$ gegeben. Unter den Voraussetzungen von Lemma 3.8, (3.1.4)*,* (3.1.10)*,* (3.1.13) *und* (3.1.14) *liegt der Speicheraufwand für eine vollständige $\mathcal{RH}^2$-Matrix-Darstellung bei*

$$(\mathcal{C}_{nuf} + 2\mathcal{C}_{cb})\left(k\#\mathcal{I} + k^2(\kappa^2(p_\mathcal{I}+1)+1)\right). \tag{3.1.15}$$

Beweis: Der Beweis erfolgt durch Addieren der Abschätzung für das Nah- und Fernfeld mit dem zweifachen Aufwand für die Clusterbasis und Zusammenfassen der Terme mit der Baumtiefe

$$\begin{aligned} &k\#\mathcal{I}(\mathcal{C}_{nuf} + 2\mathcal{C}_{cb}) + k^2\left(\kappa^2(p_\mathcal{I}+1)2\mathcal{C}_{cb} + \mathcal{C}_{nuf}(p_\mathcal{I}\kappa^2+1)\right) \\ &\quad\leq k\#\mathcal{I}(\mathcal{C}_{nuf} + 2\mathcal{C}_{cb}) + k^2(\kappa^2(p_\mathcal{I}+1)+1)(2\mathcal{C}_{cb} + C_{nuf}). \end{aligned}$$

□

3.2 Algorithmen

Ein großer Teil der Effizienz hierarchischer Matrizen beruht auf ihrer Struktur, die rekursive Algorithmen nicht nur erlaubt, sondern förmlich fordert. Dabei können zwei grundsätzliche Typen von rekursiven Algorithmen, die in dieser Arbeit Anwendung finden, unterschieden werden. Zum besseren Verständnis der folgenden Algorithmen sollen beide Ablaufschemata kurz vorgestellt werden.

Die Abbildung 3.9 zeigt schematisch, wie ein Algorithmus durch einen Clusterbaum laufen kann, hier durch die Cluster A, B, C, D, E, F und G dargestellt.

Ein *top-down*-Algorithmus beginnt zunächst beim Wurzelcluster A seine Anweisungen abzuarbeiten und fährt dann mit dem Cluster B fort. Die Anweisungen werden beim ersten

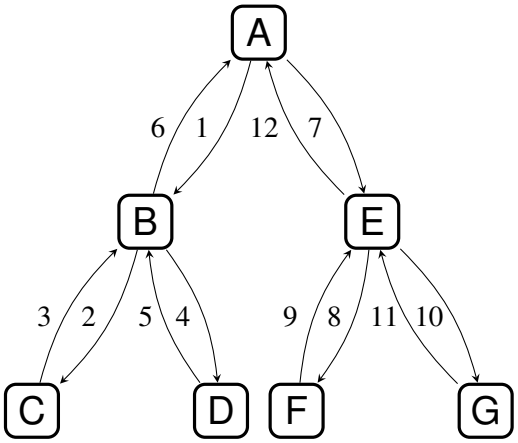


Abbildung 3.9: Ablaufschema eines Algorithmus für hierarchische Strukturen

Erreichen des entsprechenden Clusters durchgeführt. Kennzeichne das Abarbeiten der Anweisungen für alle Richtungen durch $[c]$, dann ist für die linke Hälfte des Beispielbaums folgendes Ablaufschema möglich

$$A[c] \xrightarrow{1} B[c] \xrightarrow{2} C[c] \xrightarrow{3} B \xrightarrow{4} D[c] \xrightarrow{5} B \xrightarrow{6} A \xrightarrow{7} \dots \ .$$

Hingegen werden bei einem *bottom-up*-Algorithmus die Anweisungen beim letzten Erreichen des Clusters durchgeführt. Demnach ist ein mögliches Ablaufschema für die linke Hälfte des Beispielbaums durch

$$A \xrightarrow{1} B \xrightarrow{2} C[c] \xrightarrow{3} B \xrightarrow{4} D[c] \xrightarrow{5} B[c] \xrightarrow{6} A \xrightarrow{7} \dots$$

gegeben.
Natürlich lassen sich auch beide Grundtypen zusammenführen mit einer Anweisung für den *top-down*-Teil $[c_{pre}]$ und einer für den *bottom-up* $[c_{post}]$.
Oftmals, zum Beispiel beim Berechnen der Einträge, sind die Anweisungen für die Blätter andere als für den Rest des Baums. Da Rechenoperationen nicht für die gesamte Matrix auf einmal, sondern nur für einzelne Teilmatrizen durchgeführt werden, lassen sich Algorithmen für die hierarchische Matrizen gut parallelisieren und damit beschleunigen.
Mit diesem Wissen soll einer der wichtigsten Algorithmen betrachtet werden.

3.2.1 Matrix-Vektor-Multiplikation

Eine der wichtigsten Rechenoperationen ist die Matrix-Vektor-Multiplikation, da sie für das Lösen von linearen Gleichungssystemen essentiell ist. Entsprechend ist es wichtig, dass die Rechenoperation

$$y \leftarrow y + Ax \qquad \text{für } A \in \mathbb{C}^{\mathcal{I}\times\mathcal{I}} \text{ und } x, y \in \mathbb{C}^{\mathcal{I}}$$

effizient ausgeführt werden kann [3, Kap. A].
Da bei $\mathcal{RH}^2$-Matrizen $A \in \mathbb{C}^{\mathcal{I}\times\mathcal{I}}$ nicht mehr direkt vorliegt, ist zu Beginn nicht klar, wie eine Multiplikation der Matrix mit einem Vektor $x \in \mathbb{C}^{\mathcal{I}}$ effizient durchgeführt werden kann. Jedes Blatt $b \in \mathcal{L}_{\mathcal{I}\times\mathcal{I}}$ einzeln zu behandeln, ist nicht effizient, da für zulässige Blätter mindestens drei Matrix-Vektor-Multiplikationen nötig wären. Ein schneller Algorithmus sollte zudem die vorliegende hierarchische Struktur geschickt verwenden.
Betrachte zunächst den Fall eines zulässigen Blocks $b = (t, s, c) \in \mathcal{L}^+_{\mathcal{I}\times\mathcal{I}}$, die Teilmatrix $A|_{t\times s}$ ist dann durch

$$A|_{t\times s} = V_{tc} S_b W^*_{sc}$$

gegeben, wobei die Matrizen der Clusterbasen eventuell noch rekonstruiert werden müssen. Um das Problem der wiederholten Rekonstruktion von Matrizen der Spaltenclusterbasis elegant zu umgehen, wird für jede effektive Richtung zuerst ein Vektor mit den Matrizen der Spaltenclusterbasis multipliziert. Auf diese Weise sind die Vektoren für die Multiplikation mit den Kopplungsmatrizen schon vorberechnet. Ebenso werden alle Multiplikationen mit der Zeilenclusterbasis in einem Schritt zusammengefasst. In Folge ergibt sich eine dreiphasige Vorgehensweise.
Beim ersten Schritt, der *Vorwärtstransformation*, wird der Vektor entsprechend der hierarchischen Struktur in Teilvektoren aufgeteilt und die Teilvektoren mit den Matrizen der Spaltenclusterbasis multipliziert. In den Blättern ist der erste Schritt noch offensichtlich, da $\mathrm{kind}(s) = \emptyset$ gilt, liegt W_{sc} direkt vor. Entsprechend wird in den Blättern gestartet. Die Vorwärtstransformation ist demnach ein bottom-up-Algorithmus, bei dem zunächst mit dem Hilfsvektor $\widetilde{x}_{sc} \in \mathbb{C}^{k_{sc}}$ die Multiplikation

$$\widetilde{x}_{sc} \leftarrow W^*_{sc} x|_s$$

durchgeführt wird. Handelt es sich bei dem Spaltencluster s um kein Blatt, ist die Matrix der Clusterbasis für $c' = \mathrm{r}_s(c)$ indirekt durch

$$W_{sc} = \sum_{s'\in\mathrm{kind}(s)} W_{s'c'} E_{s'c}$$

gegeben. Nachdem die Vorwärtstransformation für alle Kinder in $\mathrm{kind}(s)$ durchgeführt wurde, kann für den Cluster s und die Richtung c mit

$$W^*_{sc} x|_s = \sum_{s'\in\mathrm{kind}(s)} E^*_{s'c} W^*_{s'c'} x|_{s'} = \sum_{s'\in\mathrm{kind}(s)} E^*_{s'c} \widetilde{x}_{s'c'}$$

gearbeitet werden, was zu

$$\widetilde{x}_{sc} \leftarrow \sum_{s'\in\mathrm{kind}(s)} E^*_{s'c} \widetilde{x}_{s'c'}$$

procedure forward_transformation($\mathcal{T}_\mathcal{I}, \mathcal{R}, s, W, x, \widetilde{x}$)
  if $\operatorname{kind}(s) \neq \emptyset$ **then**
    $\widetilde{x}_{sc} \leftarrow 0$ for all $c \in \mathcal{R}_s$
    for all $s' \in \operatorname{kind}(s)$ **do**
      forward_transformation($\mathcal{T}_\mathcal{I}, \mathcal{R}, s', W, x, \widetilde{x}$)
      for all $c \in \mathcal{R}_s$ **do**
        $\widetilde{x}_{sc} \leftarrow \widetilde{x}_{sc} + E^*_{s'c}\widetilde{x}_{s'c'}$ with $c' = \mathrm{r}_s(c)$
      end for
    end for
  else
    for all $c \in \mathcal{R}_s$ **do**
      $\widetilde{x}_{sc} \leftarrow W^*_{sc} x|_s$
    end for
  end if
end procedure

Algorithmus 3.1: Die Vorwärtstransformation

führt, und damit zu einem rekursiven Algorithmus.

Da der Aufwand für eine Matrix-Vektor-Multiplikation mit einer $\mathcal{RH}^2$-Matrix ein wichtiges Kriterium ist, um den Vorteil gegenüber der vollbesetzten Matrix zu quantifizieren, wird die erste Phase direkt abgeschätzt.

Lemma 3.21 (Aufwand Vorwärtstransformation)
Der Aufwand für die Vorwärtstransformation einer Clusterbasis $\{W_{sc}\}_{\substack{s\in\mathcal{T}_\mathcal{I}\\ c\in\mathcal{R}_s}}$ zu einem gegebenen richtungsabhängigen Clusterbaum $\mathcal{T}_\mathcal{I}$, der die Voraussetzungen von Lemma 3.19 erfüllt, ist beschränkt durch

$$2k\mathcal{C}_{cb}\left(\#\mathcal{I} + k\kappa^2(p_\mathcal{I}+1)\right),$$

wobei $\mathcal{C}_{cb}$ wie in Lemma 3.19 gegeben ist.

Beweis: Jede Matrix der Spaltenclusterbasis wird nur ein einziges Mal und jeder Eintrag einer Matrix wird für maximal zwei Rechenoperationen verwendet. Damit erfordert die Multiplikation mit einem Vektor sowie die möglicherweise anschließende Addition zweimal den Aufwand für das Speichern jeder Matrix. □

Anschließend erfolgt der *Kopplungsschritt*, bei dem die Multiplikation mit der Kopplungsmatrix durchgeführt wird. Für $b = (t, s, c) \in \mathcal{L}^+_{\mathcal{I}\times\mathcal{I}}$ wird das Ergebnis der Multiplikation in einem Vektor $\widetilde{y}_{tc} \in \mathbb{C}^{k_{tc}}$ gespeichert, beziehungsweise falls die Kombination aus t und c

schon aufgetreten ist, wird zu den schon bestehenden Einträgen im Vektor addiert

$$\widetilde{y}_{tc} \leftarrow \widetilde{y}_{tc} + S_b \widetilde{x}_{sc}.$$

In diesem Schritt kann auch das Nahfeld abgearbeitet werden. Hierbei werden dann direkt der Ausgangs- und Zielvektor verwendet, entsprechend wird

$$y|_{\star(t)} \leftarrow y|_{\star(t)} + N|_b x|_{\star(s)}$$

berechnet, beziehungsweise zu $y|_{\star(t)}$ addiert.

---

procedure coupling($\mathcal{T}_{\mathcal{I}\times\mathcal{I}}, b, S, N, y, x, \widetilde{y}, \widetilde{x}$)
  if $\operatorname{kind}(b) \neq \emptyset$ **then**
    for all $b' \in \operatorname{kind}(b)$ **do**
      coupling($\mathcal{T}_{\mathcal{I}\times\mathcal{I}}, b', S, N, y, x, \widetilde{y}, \widetilde{x}$)
    end for
  else
    if $b = (t, s, c) \in \mathcal{L}^+_{\mathcal{I}\times\mathcal{I}}$ **then**
      $\widetilde{y}_{tc} \leftarrow \widetilde{y}_{tc} + S_b \widetilde{x}_{sc}$
    else
      $y|_{\star(t)} \leftarrow y|_{\star(t)} + N|_b x|_{\star(s)}$
    end if
  end if
end procedure

Algorithmus 3.2: Der Kopplungsschritt

Lemma 3.22 (Aufwand Kopplungsschritt)
Für einen gegebenen richtungsabhängigen Blockbaum $\mathcal{T}_{\mathcal{I}\times\mathcal{I}}$, der die Voraussetzungen von Lemma 3.18 erfüllt, ist der Aufwand zur Berechnung der Matrix-Vektor-Multiplikation im Kopplungsschritt beschränkt durch

$$2\mathcal{C}_{nuf}(k\#\mathcal{I} + k^2(p_{\mathcal{I}}\kappa^2 + 1)),$$

wobei $\mathcal{C}_{nuf}$ wie in Lemma 3.18 gegeben ist.

Beweis: Auch im Kopplungsschritt werden alle Matrizen des Nah- und Fernfelds nur ein einziges Mal verwendet, so dass die Aufwandsabschätzung direkt aus dem Resultat für den benötigten Speicher folgt. □

Es fehlt noch die Multiplikation mit den Matrizen der Zeilenclusterbasis sowie das Zusammenfügen des Zielvektors. Die Multiplikation mit den Matrizen der Zeilenclusterbasis erfolgt als top-down-Algorithmus und wird als *Rückwärtstransformation* bezeichnet. Für

ein $t \in \mathcal{T}_\mathcal{I}$ mit $\operatorname{kind}(t) \neq \emptyset$ wird entsprechend für alle seine Kinder $t' \in \operatorname{kind}(t)$ der Vektor $\widetilde{y}_{tc}$ auf t' und $c' = \mathrm{r}_t(c)$ transformiert und zu den dort möglicherweise schon vorhandenen Vektoreinträgen addiert

$$\widetilde{y}_{t'c'} \leftarrow \widetilde{y}_{t'c'} + E_{t'c}\widetilde{y}_{tc}.$$

Im Fall eines Blattclusters t kann die Matrix V_{tc} direkt genutzt werden

$$y|_t \leftarrow y|_t + V_{tc}\widetilde{y}_{tc}.$$

---

procedure backward_transformation($\mathcal{T}_\mathcal{I}, \mathcal{R}, t, V, y, \widetilde{y}$)
  if $\operatorname{kind}(t) \neq \emptyset$ **then**
    for all $t' \in \operatorname{kind}(t)$ **do**
      for all $c \in \mathcal{R}_t$ **do**
        $\widetilde{y}_{t'c'} \leftarrow \widetilde{y}_{t'c'} + E_{t'c}\widetilde{y}_{tc}$ with $c' = \mathrm{r}_t(c)$
      end for
      backward_transformation($\mathcal{T}_\mathcal{I}, \mathcal{R}, t', V, y, \widetilde{y}$)
    end for
  else
    for all $c \in \mathcal{R}_t$ **do**
      $y|_t \leftarrow y|_t + V_{tc}\widetilde{y}_{tc}$
    end for
  end if
end procedure

Algorithmus 3.3: Die Rückwärtstransformation

Auch der Aufwand für die Rückwärtstransformation lässt sich durch den Speicherbedarf beschränken.

Lemma 3.23 (Aufwand Rückwärtstransformation)
Der Aufwand für die Rückwärtstransformation einer Clusterbasis $\{V_{tc}\}_{\substack{t\in\mathcal{T}_\mathcal{I}\\ c\in\mathcal{R}_t}}$ zu einem gegebenen richtungsabhängigen Clusterbaum $\mathcal{T}_\mathcal{I}$, der die Voraussetzungen von Lemma 3.19 erfüllt, ist durch

$$2k\mathcal{C}_{cb}\left(\#\mathcal{I} + k\kappa^2(p_\mathcal{I}+1)\right),$$

beschränkt, dabei ist $\mathcal{C}_{cb}$ wie in Lemma 3.19 gegeben.

Beweis: Auch dieser Beweis folgt direkt aus der Speicheraufwandsabschätzung für richtungsabhängige Clusterbasen. □

Die Matrix-Vektor-Multiplikation ist mit einem Aufwand von $\mathcal{O}(\#\mathcal{I}k + k^2\kappa^2 \log_2(\#\mathcal{I}))$ (unter Verwendung von Bemerkung 12) statt $\mathcal{O}((\#\mathcal{I})^2)$ für eine vollbesetzte Matrix durchführbar. Besonders bei großen Matrizen und iterativen Verfahren, die viele Matrix-Vektor-Multiplikationen benötigen, kann die Ersparnis enorm sein.

3.3 Numerische Experimente

In diesem Abschnitt sollen Ergebnisse der $\mathcal{RH}^2$-Matrix-Approximation aus der Programmbibliothek *H2Lib*[iii] gezeigt werden.

3.3.1 Aufwand

Zunächst soll ein kleines Experiment den Unterschied der Menge der theoretisch möglichen und effektiven Richtungen zeigen. Dazu wurde ein richtungsabhängiger Blockbaum für die Sphäre mit $\#\mathcal{I} = 32768$ und der Wellenzahl $\kappa = 32$ aufgestellt. Die Auflösung betrug 32 und die Zulässigkeitsparameter wurden mit $\eta_1 = 10$ und $\eta_2 = 1$ gewählt. Nach dem Erstellen des Blockbaums wurden nicht benötigte Richtungen aussortiert, so dass nur die nötigen Richtungen auf Stufe $\ell \in \underline{p_{\mathcal{I}\rfloor 0}}$ in $\mathcal{R}_\ell^{neu}$ enthalten sind. Die Ergebnisse finden sich in der Tabelle 3.1 und zeigen, dass die Zahl der effektiven Richtungen eines Clusters deutlich von der Zahl der möglichen abweichen kann. Dass der Algorithmus bis Stufe sieben alle Richtungen bis auf eine aussortiert hat, liegt daran, dass bis dahin kein Block zulässig ist und Richtungen für unzulässige Blöcke in der Praxis keine Rolle spielen. Auf Stufe acht und neun werden nicht alle möglichen Richtungen verwendet, so dass für Cluster auf diesen beiden Stufen definitiv $\#\mathcal{R}_t > \#\mathcal{R}_t^{eff}$ gilt. Auf den Stufen zehn und elf werden alle möglichen Richtungen zumindest einmal bei einem Block verwendet, für Cluster auf diesen beiden Stufen gilt daher $\#\mathcal{R}_t \geq \#\mathcal{R}_t^{eff}$. Auf Stufe 12 ist dann nur noch die Null-Richtung vorhanden.

Im Folgenden wird der Speicheraufwand der $\mathcal{RH}^2$-Matrix in Bezug auf Veränderung in den Parametern $\#\mathcal{I}$ und κ untersucht. Da an diesem Punkt noch kein Interesse am Approximationsfehler besteht, beschränkt sich das Experiment allein auf den Einfachschichtoperator, denn in puncto richtungsabhängiger Blockbaum und Speicheraufwand gibt es keinen Unterschied zum Doppelschichtoperator. Um die Ergebnisse realistisch zu gestalten, sind die Parameter, wenn nicht anders angegeben, mit $m = 3$ für die Interpolationsordnung (da der Einfluss der Interpolationsordnung auf den Speicher klar ist, bleibt sie fest) und

[iii] Die Programmbibliothek wird von der Arbeitsgruppe Scientific Computing der Universität zu Kiel herausgegeben und Teile des Quellcodes wurden im Zuge dieser Arbeit selbst geschrieben. Die aktuell erhältliche Version ist unter http://www.h2lib.org/ zu finden.

Stufe ℓ	$\#\mathcal{R}_\ell$	$\#\mathcal{R}_\ell^{neu}$		Stufe ℓ	$\#\mathcal{R}_\ell$	$\#\mathcal{R}_\ell^{neu}$
0	1536	1		7	96	1
1	1176	1		8	54	8
2	864	1		9	24	16
3	384	1		10	24	24
4	294	1		11	24	24
5	216	1		12	1	1
6	96	1				

Tabelle 3.1: Anzahl der möglichen und verwendeten Richtungen

$res = 2 * (m+1)^3 = 128$ für die Auflösung (siehe Bemerkung 17) gegeben. Die Zulässigkeitsparameter wurden mit $\eta_1 = 10$ und $\eta_2 = 1$ gewählt, lediglich die Wellenzahl wurde angepasst zur Anzahl der Freiheitsgrade gewählt, so dass im hochfrequenten Fall $\kappa * h \approx 1,22$ (im niedrigfrequenten Fall $\kappa * h \approx 0,3$) für den maximalen Durchmesser h eines Triangulationselements gewährleistet ist.
Zunächst betrachte einen Datensatz zur Abhängigkeit des Speicheraufwands der $\mathcal{RH}^2$-Matrix ($\widetilde{A}_e$) von der Mächtigkeit der Indexmenge $n := \#\mathcal{I}$. Die Speicheranforderung der vollbesetzten Matrix (A_e) kann durch die Formel $16n^2$ Bytes[iv] bestimmt werden.

Die Tabellen 3.2 und 3.3 zeigen Vergleichswerte für den Speicheraufwand im hochfrequenten (Tab. 3.2) und niedrigfrequenten (Tab. 3.3) Fall. In der Spalte für die $\mathcal{RH}^2$-Matrix ($\widetilde{A}_e$) ist der Speicher für die komplette Matrix, also inklusive zweier Clusterbasen, aber ohne Blockbaum angegeben. In den runden Klammer befindet sich die jeweilige Speicheranforderung pro Freiheitsgrad.
Höhere Wellenzahlen implizieren schärfere Zulässigkeitsbedingungen und sorgen somit für vollere Blockbäume sowie einen höheren Nahfeldanteil, was sich im Vergleich der beiden Datensätze widerspiegelt.

Die Ergebnisse der beiden Tabellen sind noch einmal in der Abbildung 3.10 grafisch dargestellt. Während der Aufwand für die vollbesetzte Matrix entsprechend des Darstellungsformates linear wächst, zeigen die Approximationen ein anderes Verhalten.
Der angeforderte Speicher deutet im niedrigfrequenten Fall auf ein Verhalten mit $\mathcal{O}(kn)$ hin, was dem typischen Verhalten einer $\mathcal{H}^2$-Matrix entspricht. Das theoretisch zu erwartende $\mathcal{O}(kn + k^2\kappa^2 \log_2(n))$ beschreibt das Verhalten im hochfrequenten Fall besser. Das asymptotische Verhalten lässt sich hier jedoch erst ziemlich spät erkennen.

[iv] Die 16 ergibt sich durch die Verwendung doppelter Genauigkeit (8B) für komplexe Zahlen ($2 \cdot 8$B).

n	κ	$\widetilde{A}_e$ [MiB]	([KiB]/n)	A_e [MiB]	([KiB]/n)
2 048	8	64.2	(32.0)	64.0	(32.0)
4 608	12	324.8	(72.2)	324.0	(72.0)
8 192	16	1 025.9	(128.3)	1 024.0	(128.0)
18 432	24	5 191.1	(288.4)	5 184.0	(288.0)
32 768	32	16 404.2	(512.6)	16 384.0	(512.0)
73 728	48	81 656.3	(1 134.2)	82 944.0	(1 152.0)
131 072	64	243 736.1	(1 904.2)	262 144.0	(2 048.0)
294 912	96	1 063 580.9	(3 692.9)	1 327 104.0	(4 608.0)

Tabelle 3.2: Speicheraufwand bei hohen Wellenzahlen

n	κ	$\widetilde{A}_e$ [MiB]	([KiB]/n)	A_e [MiB]	([KiB]/n)
2 048	2	64.0	(32.0)	64.0	(32.0)
4 608	3	323.3	(71.9)	324.0	(72.0)
8 192	4	1 007.1	(125.9)	1 024.0	(128.0)
18 432	6	4 402.2	(244.6)	5 184.0	(288.0)
32 768	8	10 874.3	(339.8)	16 384.0	(512.0)
73 728	12	37 423.4	(519.8)	82 944.0	(1 152.0)
131 072	16	82 908.0	(647.8)	262 144.0	(2 048.0)
294 912	24	229 113.7	(795.6)	1 327 104.0	(4 608.0)

Tabelle 3.3: Speicheraufwand bei niedrigen Wellenzahlen

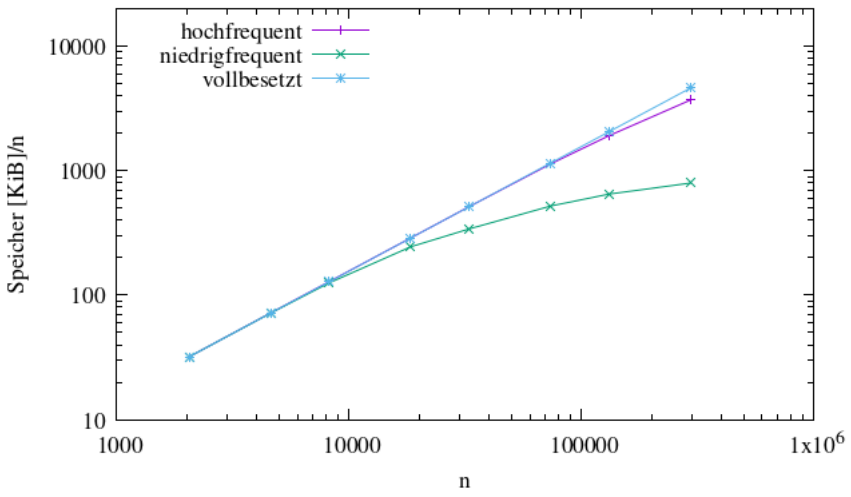


Abbildung 3.10: Grafische Darstellung des Speicheraufwands

Auch in Bezug auf die Wellenzahl verhält sich die $\mathcal{RH}^2$-Matrix wie erwartet, höhere Wellenzahlen sorgen für eine aufwendigere Approximation, wie auch Tabelle 3.4 zeigt. Zusätzlich ist hier die Anzahl der Blöcke des richtungsabhängigen Blockbaums angegeben.

In Tabelle 3.5 finden sich Ergebnisse zum zeitlichen Aufwand der Berechnungen. Erneut wurden für unterschiedliche Problemgrößen ein hoch- sowie ein niedrigfrequenter Fall betrachtet. Die Zeitberechnungen wurden auf einem 32-Kern-Computer mit Intel® Xeon® CPU E7-4809 Prozessoren und einer Taktfrequenz von 2.00GHz durchgeführt, bei der zusätzlich ein parallelisierter Code verwendet wurde. Es zeigt sich, dass in beiden Fällen eine Zeitersparnis im Aufstellen der $\mathcal{RH}^2$-Matrix gegenüber der vollbesetzten Matrix zu beobachten ist, auch wenn diese im Falle von höheren Wellenzahlen geringer ausfällt.
Auffällig ist, dass die Approximation im hochfrequenten Bereich noch immer sehr speicherintensiv ist. Entsprechend muss es Ziel sein, die $\mathcal{RH}^2$-Matrix auch im hochfrequenten Bereich speichereffizienter zu gestalten, was Thema der Kapitel 5 und 6 sein wird.

Blöcke	κ	$\widetilde{A}_e$ [MiB]	([KiB]/n)
151 477	1	12 301.8	(96.1)
158 005	2	12 606.4	(98.5)
227 493	4	16 128.2	(125.9)
482 997	8	33 849.6	(264.5)
1 070 357	16	82 908.0	(647.8)
1 689 301	32	163 728.8	(1 279.1)
2 080 213	64	243 736.1	(1 904.2)

Tabelle 3.4: Anzahl der Blöcke bei verschiedenen Wellenzahlen für $n = 131\,072$

	niedrigfrequent			hochfrequent		
n	κ	$\widetilde{A}_e$ [s]	A_e [s]	κ	$\widetilde{A}_e$ [s]	A_e [s]
2 048	2	5.1	5.4	8	5.7	5.4
4 608	3	14.4	20.7	12	14.9	22.5
8 192	4	39.4	61.5	16	43.4	60.9
18 432	6	150.1	284.6	24	211.8	294.6
32 768	8	360.0	690.2	32	657,3	674.9
73 728	12	1 216.0	3 114.5	48	3 164.4	3 239.5
131 072	16	2 129.8	9 850.8	64	8 976.3	9 986.3

Tabelle 3.5: Benötigte Zeit zum Bestimmen der Einträge

3.3.2 Matrix-Vektor-Multiplikation

Die Zeitberechnungen für die Matrix-Vektor-Multiplikation wurden auf einem *Shared Memory* System mit zwei Intel® Xeon® Platinum 8160 Prozessoren mit insgesamt 48 Kernen durchgeführt, der Code für die Matrix-Vektor-Multiplikation wurde nicht parallelisiert. Um einen brauchbaren Mittelwert für die benötigte Zeit zu erhalten, wurde die Berechnung fünfzig mal direkt hintereinander durchgeführt und die verwendeten Vektoren wurden zufällig belegt. Es wurde einmal der niedrig- und einmal der hochfrequente Fall betrachtet. Die Ergebnisse des Experiments finden sich in Tabelle 3.6, dabei zeigt die erste Spalte die Anzahl der Freiheitsgrade $n := \#\mathcal{I}$, in Spalte zwei und fünf sind die Wellenzahlen zu finden. In den Spalten drei und sechs ist die minimal aufgetretene Laufzeit vermerkt, während sich der jeweilige Mittelwert in Spalte vier und sieben befindet.

Nach Kapitel 3.2.1 ist die Komplexität der Matrix-Vektor-Multiplikation ebenfalls durch $\mathcal{O}(kn + k^2\kappa^2 \log_2(n))$ beschränkt. Der Tabelle 3.6 ist zu entnehmen, dass die benötigte Zeit für die Matrix-Vektor-Multiplikation genau wie der Speicheraufwand (siehe Tab. 3.4) mit einer wachsenden Wellenzahl steigt.
Die Abbildung 3.11 visualisiert die Ergebnisse für die durchschnittliche Zeit aus Tabelle 3.6 noch einmal, wobei zusätzlich Vergleichskurven des erwarteten Aufwands in blau beziehungsweise gelb eingezeichnet sind. In beiden Fällen ist die Abschätzung am Anfang zu grob und der Bereich, in dem das asymptotische Verhalten deutlich hervor tritt, scheint noch nicht erreicht.

Dass die Matrix-Vektor-Multiplikation trotz ähnlicher Aufwandsabschätzung langsamer zu sein scheint, liegt wahrscheinlich daran, dass die Matrizen zu klein sind, so dass die Speicherzugriffe den Aufwand dominieren. Diese Vermutung wird durch das Verhältnis von Laufzeit zu Speicher gestützt, welches nahezu konstant und in Tabelle 3.7 zu finden ist.

	niedrigfrequent			hochfrequent		
n	κ	min [s]	∅ [s]	κ	min [s]	∅ [s]
2048	2	0.0	0.0	8	0.0	0.0
4608	3	0.1	0.1	12	0.1	0.1
8192	4	0.2	0.23	16	0.2	0.24
18432	6	1.0	1.23	24	1.2	1.42
32768	8	2.5	2.79	32	3.8	4.44
73728	12	8.9	10.28	48	20.0	26.55
131072	16	19.6	26.83	64	105.2	106.49

Tabelle 3.6: Laufzeit Matrix-Vektor-Multiplikation

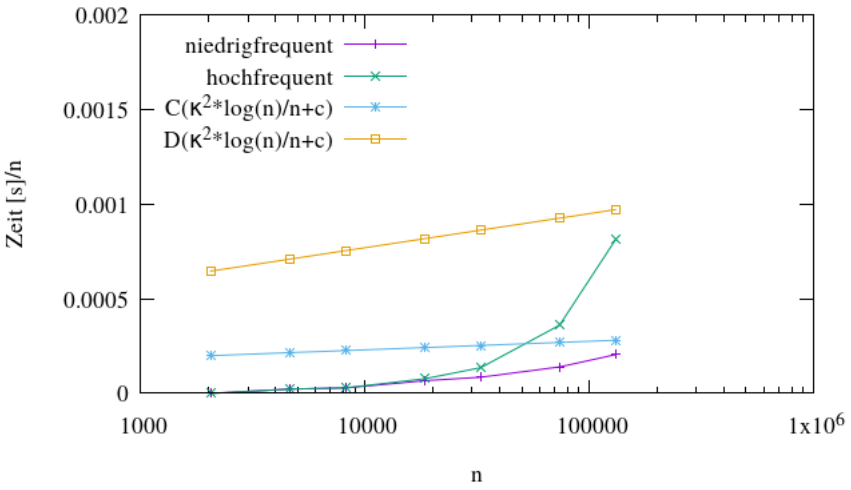


Abbildung 3.11: Benötigte Zeit für die Matrix-Vektor-Multiplikation

	niedrigfrequent		hochfrequent	
n	κ	Zeit/Speicher [s/KiB]	κ	Zeit/Speicher [s/KiB]
4608	3	0.308_{-6}	12	0.307_{-6}
8192	4	0.227_{-6}	16	0.228_{-6}
18432	6	0.273_{-6}	24	0.267_{-6}
32768	8	0.251_{-6}	32	0.264_{-6}
73728	12	0.269_{-6}	48	0.318_{-6}
131072	16	0.316_{-6}	64	0.427_{-6}

Tabelle 3.7: Verhältnis Laufzeit zu Speicher

4 Fehlerabschätzungen

Wichtig ist nicht nur die speichergünstige Darstellung, sondern auch, dass eine benötigte Genauigkeit in der Matrixapproximation erreicht werden kann. In diesem Kapitel wird wie in [3] gezeigt, dass theoretisch jede gewünschte Genauigkeit möglich ist und der Approximationsfehler exponentiell konvergiert.

4.1 Erweiterte Interpolationsfehlerabschätzungen

Typische Aussagen zur Approximationsgüte der Interpolation nutzen Ableitungen der Funktion oder ein Bestapproximationsargument. Da weder der Fehler der Bestapproximation bekannt ist, noch höhere Ableitungen der Funktion vorliegen oder leicht zu bestimmen sind, muss ein längerer Weg zur Fehlerabschätzung genommen werden.
Für holomorphe Funktionen existiert eine Quasi-Bestapproximationsaussage, welche wiederum die sogenannten *Bernstein-Ellipsen*[i] als Gebiete benötigt [32, S. 364 ff.]. Die Bernstein-Ellipsen ersetzen in der Theorie das Intervall $[-1, 1]$ und nähern sich im Grenzfall dem Einheitsintervall an.

Definition 4.1 (Bernstein-Ellipse)
Die Bernstein-Ellipse *zu einem Parameter $\varrho \in \mathbb{R}_{>1}$ hat die Halbachsen*

$$a_\varrho = \frac{\varrho+\frac{1}{\varrho}}{2}, \qquad b_\varrho = \frac{\varrho-\frac{1}{\varrho}}{2}$$

und ihr Inneres ist entsprechend durch

$$D_\varrho := \left\{ z \in \mathbb{C} \mid \left(\frac{\Re(z)^2}{a_\varrho^2}\right) + \left(\frac{\Im(z)^2}{b_\varrho^2}\right) < 1 \right\}$$

gegeben. Bezeichne den Rand der Ellipse mit

$$\partial D_\varrho = \left\{ z \in \mathbb{C} \mid \left(\frac{\Re(z)^2}{a_\varrho^2}\right) + \left(\frac{\Im(z)^2}{b_\varrho^2}\right) = 1 \right\}.$$

Der Parameter ϱ ist gleich der Halbachsensumme der Bernstein-Ellipse, denn offensichtlich gilt $\varrho = a_\varrho + b_\varrho$. Allen Bernstein-Ellipsen ist per Konstruktion gemein, dass die Brennpunkte der Ellipsen bei $(-1, 0)$ und $(1, 0)$ liegen [32, S. 365].

[i] Benannt nach dem russischen Mathematiker Sergei Natanowitsch Bernstein, der sich unter anderem mit der Approximationstheorie beschäftigte.

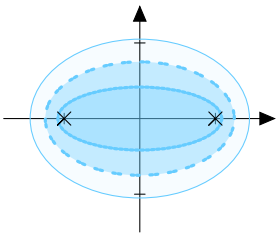


Abbildung 4.1: Bernstein-Ellipsen

Die Abbildung 4.1 zeigt Bernstein-Ellipsen für die Parameter $\varrho = \frac{3}{2}, 2, \frac{5}{2}$ (von innen nach außen).
Die Definition der Bernstein-Ellipse nutzt die Halbachsen zur Charakterisierung, eine Ellipse kann aber auch äquivalent dazu über die Brennpunkte charakterisiert werden. Das folgende Lemma zeigt die Äquivalenz der beiden Formulierungen und erweitert damit die Möglichkeiten, die Bernstein-Ellipse zu beschreiben.

Lemma 4.2
Die Bernstein-Ellipse zum Parameter $\varrho > 1$ erfüllt

$$\overline{D}_\varrho = \{z \in \mathbb{C} \mid |z+1| + |z-1| \leq 2a_\varrho\},$$

wobei a_ϱ wie in der Definition 4.1 die reelle Halbachse sei.

Beweis: Bevor die Äquivalenz der beiden charakterisierenden Ungleichungen gezeigt wird, mache noch eine kurze hilfreiche Beobachtung.
Grundsätzlich gilt für die Halbachsen a_ϱ, b_ϱ der Bernstein-Ellipse zum Parameter $\varrho > 1$ die folgende Beziehung

$$b_\varrho^2 = \left(\frac{\varrho - \frac{1}{\varrho}}{2}\right)^2 = \frac{\varrho^2 - 2 + \frac{1}{\varrho^2}}{4} = \frac{\varrho^2 + 2 + \frac{1}{\varrho^2} - 4}{4} = \left(\frac{\varrho + \frac{1}{\varrho}}{2}\right)^2 - 1 = a_\varrho^2 - 1. \tag{4.1.1}$$

Betrachte einen Punkt $z \in \mathbb{C}$ mit $z = x + \mathrm{i}y$ für $x, y \in \mathbb{R}$, der

$$|z+1| + |z-1| \leq 2a_\varrho$$

erfüllt. Verwende die Definition des Betrags einer komplexen Zahl, um die linke Seite der Ungleichung umzuformen, und sortiere um

$$\begin{aligned} &\sqrt{(x+1)^2 + y^2} + \sqrt{(x-1)^2 + y^2} \leq 2a_\varrho \iff \\ &\sqrt{(x-1)^2 + y^2} \leq 2a_\varrho - \sqrt{(x+1)^2 + y^2}. \end{aligned}$$

Quadrieren, Ausmultiplizieren und Weglassen von Termen, die auf beiden Seiten auftauchen, führt zu

$$\begin{aligned}(x-1)^2+y^2 &\leq \left(2a_\varrho - \sqrt{(x+1)^2+y^2}\right)^2 \iff\\ (x-1)^2+y^2 &\leq 4a_\varrho^2 - 4a_\varrho\sqrt{(x+1)^2+y^2} + (x+1)^2 + y^2 \iff\\ -2x+x^2+1+y^2 &\leq 4a_\varrho^2 - 4a_\varrho\sqrt{(x+1)^2+y^2} + 2x + x^2 + 1 + y^2 \iff\\ -4x &\leq 4a_\varrho^2 - 4a_\varrho\sqrt{(x+1)^2+y^2}.\end{aligned}$$

Multipliziere mit $\frac{1}{4}$, stelle erneut um und quadriere danach, um

$$a_\varrho\sqrt{(x+1)^2+y^2} \leq a_\varrho^2 + x \iff a_\varrho^2\left((x+1)^2+y^2\right) \leq (a_\varrho^2+x)^2$$

zu erhalten. Ausmultiplizieren der binomischen Formeln und Umstellen führt zu

$$\begin{aligned}a_\varrho^2(x^2+y^2+2x+1) &\leq (a_\varrho^2+x)^2 \iff\\ x^2a_\varrho^2 + y^2a_\varrho^2 + 2xa_\varrho^2 + a_\varrho^2 &\leq a_\varrho^4 + x^2 + 2xa_\varrho^2 \iff\\ x^2a_\varrho^2 - x^2 + y^2a_\varrho^2 + 2xa_\varrho^2 &\leq a_\varrho^4 - a_\varrho^2 + 2xa_\varrho^2.\end{aligned}$$

Der Term $2xa_\varrho^2$ tritt auf beiden Seiten der Ungleichung auf und kann entsprechend weggelassen werden

$$x^2a_\varrho^2 - x^2 + y^2a_\varrho^2 \leq a_\varrho^4 - a_\varrho^2 \iff x^2(a_\varrho^2-1) + y^2a_\varrho^2 \leq a_\varrho^2(a_\varrho^2-1).$$

Nutze nun die Beziehung der Halbachsen (4.1.1) der Ellipse zu einander, um

$$x^2b_\varrho^2 + y^2a_\varrho^2 \leq a_\varrho^2b_\varrho^2$$

zu erhalten. Eine Multiplikation mit dem positiven Bruch $\frac{1}{a_\varrho^2b_\varrho^2}$ führt dann zur bestimmenden Ungleichung der Bernstein-Ellipse

$$\tfrac{x^2}{a_\varrho^2} + \tfrac{y^2}{b_\varrho^2} \leq 1.$$

Da zur Überführung nur Äquivalenzumformungen verwendet und auch beim Quadrieren keine negativen Vorzeichen verloren wurden, sind die beiden Charakterisierungen äquivalent. □

Bemerkung 18 (Geschachtelte Ellipsen): *Für $\varrho > \widehat{\varrho} > 1$ gilt $a_\varrho > a_{\widehat{\varrho}}$ und damit liefert die Charakterisierung über die Brennpunkte sofort*

$$\overline{D}_{\widehat{\varrho}} \subset \overline{D}_\varrho \qquad \text{für alle } \varrho > \widehat{\varrho} > 1.$$

Zusätzlich ist es möglich, die Bernstein-Ellipse über die *Joukowski-Transformation*[ii] bijektiv mit Kreisen aus der komplexen Ebene zu identifizieren [41, S. 269 ff.][20, S. 99 f.]. Diese zweite Möglichkeit der Identifikation erleichtert es, Aussagen über die Bernstein-Ellipsen zu zeigen.

Definition 4.3 (Joukowski-Transformation)
Die Abbildung

$$\varpi : \mathbb{C} \setminus \{0\} \to \mathbb{C}, \qquad z \mapsto \tfrac{1}{2}\left(z + \tfrac{1}{z}\right)$$

wird als Joukowski-Transformation *bezeichnet.*

Im folgenden Lemma sind einige leicht nachzurechnende Eigenschaften der Joukowski-Transformation zusammengefasst. Die Aussagen zur Joukowski-Transformation und ihr Zusammenhang zur Bernstein-Ellipse stammen aus einer unveröffentlichten Arbeit von Herrn Börm [9], einige Teile finden sich auch in einem Vorlesungsskript von ihm [8].

Lemma 4.4 (Eigenschaften Joukowski-Transformation)
Die Joukowski-Transformation weist folgende Eigenschaften auf

$$\varpi(z) = \varpi\left(\tfrac{1}{z}\right) \qquad \text{für alle } z \in \mathbb{C} \setminus \{0\}, \tag{4.1.2}$$

$$\varpi(z) = \Re(z) \qquad \text{für alle } z \in \mathbb{C} \text{ mit } |z| = 1, \tag{4.1.3}$$

$$x < y \iff \varpi(x) < \varpi(y) \qquad \text{für alle } x, y \in \mathbb{R}_{\geq 1}, \tag{4.1.4}$$

$$x = \varpi\left(x + \sqrt{x^2 - 1}\right) \qquad \text{für alle } x \in \mathbb{R}_{\geq 1}. \tag{4.1.5}$$

Beweis: Die erste Eigenschaft folgt aus der Definition der Joukowski-Transformation.
Für die zweite Eigenschaft sei $z \in \mathbb{C}$ mit $|z| = 1$. Direktes Nachrechnen mit der Joukowski-Transformation liefert

$$\begin{aligned}\varpi(z) &= \tfrac{1}{2}\left(z + \tfrac{1}{z}\right) = \tfrac{1}{2}\left(z + \tfrac{\bar{z}}{z\bar{z}}\right) = \tfrac{1}{2}\left(z + \tfrac{\bar{z}}{|z|^2}\right) = \tfrac{1}{2}(z + \bar{z}) \\ &= \tfrac{1}{2}(\Re(z) + \mathrm{i}\Im(z) + \Re(z) - \mathrm{i}\Im(z)) = \Re(z)\end{aligned}$$

und damit (4.1.3). Für die dritte Eigenschaft seien $x, y \in \mathbb{R}_{\geq 1}$ mit $\varpi(x) < \varpi(y)$ gegeben. Dann gelten folgende Äquivalenzen

$$\begin{aligned}\varpi(x) < \varpi(y) &\iff 2\varpi(x) < 2\varpi(y) \iff x + \tfrac{1}{x} < y + \tfrac{1}{y} \\ &\iff x - \tfrac{1}{y} < y - \tfrac{1}{x} \iff \left(1 - \tfrac{1}{xy}\right) x < \left(1 - \tfrac{1}{xy}\right) y \\ &\iff x < y.\end{aligned}$$

[ii] Die Joukowski-Transformation ist nach dem russischen Mathematiker und Aerodynamiker Nikolai Jegorowitsch Joukowski benannt. Sie wird in der Aerodynamik genutzt, um einen Kreis bijektiv auf das Profil einer Flügeltragfläche abzubilden.

Da es nicht möglich ist eine Umkehrabbildung explizit anzugeben, suche zu einem $x \in \mathbb{R}_{\geq 1}$ ein $z \in \mathbb{R}_{\geq 1}$, so dass $\varpi(z) = x$ gilt. Dies ist äquivalent zu

$$\varpi(z) = x \iff \tfrac{1}{2}\left(z + \tfrac{1}{z}\right) = x \iff z^2 + 1 = 2xz \iff z^2 + 1 - 2xz = 0.$$

Die Nullstellen des Polynoms können mit der quadratischen Ergänzung gefunden werden

$$z^2 + 1 - 2xz = 0 \iff (z - x)^2 - (x^2 - 1) = 0 \iff z = x \pm \sqrt{x^2 - 1}.$$

Wegen $x \in \mathbb{R}_{\geq 1}$ gilt $x^2 - 1 \geq 0$ und damit ist die Wurzel wohldefiniert und reell. Somit erfüllt $z := x + \sqrt{x^2 - 1}$ die Behauptung. □

Die Umkehrfunktion (4.1.5) kann zu einer Rechtsinversen auf $\mathbb{C}$ fortgesetzt werden.

Korollar 4.5 (Rechtsinverse)
Es existiert eine Abbildung

$$\varpi^\dagger : \mathbb{C} \to \{z \in \mathbb{C} \mid |z| \geq 1\},$$

die $\varpi(\varpi^\dagger(z)) = z$ für alle $z \in \mathbb{C}$ erfüllt.

Beweis: Zeige, dass ϖ surjektiv ist und damit, dass eine Rechtsinverse punktweise für alle $z \in \mathbb{C}$ definiert werden kann. Sei ein $z \in \mathbb{C}$ gegeben und suche dazu ein $\hat{z} \in \mathbb{C} \setminus \{0\}$ mit $\varpi(\hat{z}) = z$. Es gilt

$$\tfrac{1}{2}\left(\hat{z} + \tfrac{1}{\hat{z}}\right) = z \iff \hat{z}^2 - 2z\hat{z} + 1 = 0$$

und wie im Beweis zu (4.1.5) hat dieses Polynom über $\mathbb{C}$ zwei Lösungen $z_1, z_2 \in \mathbb{C}$. Nehme an, dass $|z_1| \geq |z_2|$, da die Lösungen $z_1 z_2 = 1$ erfüllen müssen, folgt dann $1 = |z_1||z_2| \leq |z_1|^2$ und damit $|z_1| \geq 1$. Setze $\hat{z} = z_1$ und erhalte so ein Urbild von z in der Menge $\{\hat{z} \in \mathbb{C} \mid |\hat{z}| \geq 1\}$. □

Mit diesen Eigenschaften ist es möglich, den Zusammenhang zwischen Kreis, Joukowski-Transformation und Bernstein-Ellipse aufzuzeigen [20, S.99 f.][8], der Beweis orientiert sich an einer unveröffentlichten Arbeit von Herrn Börm [9].

Lemma 4.6
Für $\varrho \in \mathbb{R}_{>1}$ bildet die Joukowski-Transformation den Kreis $\{z \in \mathbb{C} \mid |z| = \varrho\}$ bijektiv auf den Rand der Bernstein-Ellipse ∂D_ϱ ab.
Weiterhin gilt für alle $z \in \mathbb{C}$ mit $\frac{1}{\varrho} \leq |z| \leq \varrho$, dass $\varpi(z) \in \overline{D}_\varrho$, und für alle $w \in \overline{D}_\varrho$ existiert ein $z \in \mathbb{C}$ mit $1 \leq |z| \leq \varrho$ und $\varpi(z) = w$.

Beweis: Bevor es um die Bijektivität der Joukowski-Transformation auf den betrachteten Mengen gehen soll, mache zunächst eine allgemeine Beobachtung. Seien ein $z \in \mathbb{C} \setminus \{0\}$

und $x \in \mathbb{C}$ mit $x = \varpi(z)$ gegeben, dann gilt

$$\begin{aligned}|x+1| + |x-1| &= \tfrac{1}{2}|z + \tfrac{1}{z} + 2| + \tfrac{1}{2}|z + \tfrac{1}{z} - 2| \\ &= \tfrac{1}{2|z|}\left(|z^2 + 1 + 2z| + |z^2 + 1 - 2z|\right) \\ &= \tfrac{1}{2|z|}\left(|z+1|^2 + |z-1|^2\right) \\ &= \tfrac{1}{2|z|}\left((z+1)(\bar{z}+1) + (z-1)(\bar{z}-1)\right) \\ &= \tfrac{1}{2|z|}\left(2|z|^2 + 2\right) \\ &= |z| + \tfrac{1}{|z|}.\end{aligned}$$

Wenn nun zusätzlich $|z| = \varrho$ gilt, folgt unmittelbar

$$|x+1| + |x-1| = |z| + \tfrac{1}{|z|} = \varrho + \tfrac{1}{\varrho} = 2a_\varrho$$

und damit bildet die Joukowski-Transformation den betrachteten Kreis auf den Rand der Bernstein-Ellipse ab. Zeige nun, dass zu jedem Element des Rands der Bernstein-Ellipse genau ein Element des Kreises als Urbild existiert.
Sei dazu $x \in \partial D_\varrho$ gegeben, mit der Surjektivität von ϖ (siehe Beweis zu Korollar 4.5) existiert dann ein $z \in \mathbb{C}$ mit $|z| \geq 1$ und $\varpi(z) = x$. Nach der anfänglichen Betrachtung gilt für das $x \in \partial D_\varrho$ mit Urbild z

$$\varrho + \tfrac{1}{\varrho} = 2a_\varrho = |x+1| + |x-1| = |z| + \tfrac{1}{|z|}$$

und da $\varrho > 1$, folgt mit der Monotonie von ϖ (4.1.4) dann $|z| = \varrho$.
Die Lösung z ist eindeutig und die Abbildung damit auch injektiv. Denn für eine zweite Lösung $\hat{z} \in \mathbb{C}$ gilt wie im Beweis von Bemerkung 4.5 $\hat{z}z = 1$, so dass

$$|\hat{z}| = \tfrac{|z\hat{z}|}{|z|} = \tfrac{1}{|z|} = \tfrac{1}{\varrho}$$

folgt und damit ist dann $|\hat{z}| \neq \varrho$ erfüllt.
Für den zweiten Teil der Behauptung sei eine Bernstein-Ellipse zum Parameter $\varrho > 1$ gegeben, betrachte zunächst ein $z \in \mathbb{C}$ mit $1 < |z| \leq \varrho$. Setze $\widehat{\varrho} := |z|$ und sei $x = \varpi(z)$, dann gilt nach der ersten Behauptung des Lemmas, dass $x \in \partial D_{\widehat{\varrho}}$. Damit folgt

$$|x+1| + |x-1| = 2a_{\widehat{\varrho}} \leq 2a_\varrho,$$

was zu $x \in \overline{D}_\varrho$ führt. Nach (4.1.2) gilt $\varpi(z) = \varpi\left(\frac{1}{z}\right)$, womit die Behauptung entsprechend auch für $1 > |z| \geq \frac{1}{\varrho}$ folgt.
Sei nun $z \in \overline{D}_\varrho$, definiere $\alpha := \frac{|z+1|+|z-1|}{2}$, dann gilt offensichtlich $\alpha \leq a_\varrho$ und wegen

$$|z+1| + |z-1| = |z+1| + |(z+1) - 2| \geq |z+1| + 2 - |(z+1)| = 2$$

folgt $\alpha \geq 1$. Entsprechend kann (4.1.5) genutzt werden, um durch $\widehat{\varrho} := \varpi^\dagger(\alpha)$ einen Parameter für eine weitere Bernstein-Ellipse zu definieren. Nach dem ersten Teil der Behauptung

kann ein $x \in \mathbb{C}$ auf dem Rand des Kreises mit Radius $\widehat{\varrho}$ gefunden werden mit $\varpi(x) = z$. Dann gilt

$$\varpi(\widehat{\varrho}) = \alpha \leq a_\varrho = \varpi(\varrho)$$

und mit (4.1.4) folgt $|x| = \widehat{\varrho} \leq \varrho$ und damit der Rest der Behauptung. □

Entsprechend bildet die Joukowski-Transformation für ein $\varrho \in \mathbb{R}_{>1}$ den Kreisring

$$A_\varrho := \left\{ z \in \mathbb{C} \mid \tfrac{1}{\varrho} < |z| < \varrho \right\} \tag{4.1.6}$$

auf die Bernstein-Ellipse D_ϱ ab.

Bemerkung 19 (Einheitskreis und Joukowski-Transformation): *Wird der Einheitskreis* $\{z \in \mathbb{C} \mid |z| = 1\}$ *betrachtet, bildet die Joukowski-Transformation diesen wegen der Eigenschaft* (4.1.3) *auf das Referenzintervall* $[-1, 1]$ *ab. Die Rechtsinverse der Joukowski-Transformation bietet auch hier die Möglichkeit, zwischen der Betrachtung auf dem Intervall* $[-1, 1]$ *und der auf dem Einheitskreis zu wechseln.*

Um die Fehleraussagen beim Einfach- und Doppelschichtoperator auf eindimensionale Fehleraussagen zurückzuführen, ist es notwendig, die Interpolation holomorpher Funktionen auf Bernstein-Ellipsen näher zu untersuchen. Die folgenden grundsätzlichen Überlegungen dazu stammen aus einer unveröffentlichten Arbeit von Herrn Börm [8], welche auf [20, S. 229 ff.] beruht.
Der Umweg über Kreisringe erweist sich als hilfreich bei der Analyse der Interpolation von holomorphen Funktionen auf Bernstein-Ellipsen, denn auf Kreisringen kann ein weiteres Approximationspolynom definiert werden, mit dem sich dann die gewünschten Fehleraussagen herleiten lassen. Eine auf der Bernstein-Ellipse holomorphe Funktion $f : D_\varrho \to \mathbb{C}$ kann mit Hilfe der Joukowski-Transformation auch als eine holomorphe Funktion auf dem Kreisring A_ϱ über $\hat{f} = f \circ \varpi$ geschrieben werden, da die Joukowski-Transformation auf $\mathbb{C}\backslash\{0\}$ holomorph und $0 \notin A_\varrho$ ist, ist $\hat{f}$ als Verkettung holomorpher Funktionen holomorph. Auf Kreisringen holomorphe Funktionen können auch mit Hilfe der *Laurent-Reihe*[iii] dargestellt werden [36, Satz 16]

$$\hat{f}(\hat{z}) = \sum_{k=-\infty}^{\infty} c_k \hat{z}^k \qquad \text{für alle } \hat{z} \in A_\varrho,$$

wobei die Koeffizienten c_k für ein beliebiges $r \in (\varrho^{-1}, \varrho)$ durch

$$c_k := \tfrac{1}{2\mathrm{i}\pi} \int_{|s|=r} \tfrac{\hat{f}(s)}{s^{k+1}} \, \mathrm{d}s \qquad \text{für alle } k \in \mathbb{Z}$$

[iii] Namensgebend ist der französische Mathematiker Pierre Alphonse Laurent.

gegeben sind. Nach Lemma 4.4 gilt mit (4.1.2) $\varpi(\varrho) = \varpi(\varrho^{-1})$, so dass mit zwei passenden Parametrisierungen und der Cauchy-Integralformel[iv] [36, K. 2]

$$c_{-k} = \frac{1}{2\mathrm{i}\pi} \int_{|s|=\varrho^{-1}} \frac{\hat{f}(s)}{s^{k+1}}\,\mathrm{d}s = \frac{1}{2\mathrm{i}\pi} \int_{|s|=\varrho} \frac{\hat{f}(s)}{s^{k+1}}\,\mathrm{d}s = c_k$$

gezeigt werden kann [37, K. 5.10]. Dies ermöglicht es, die Laurent-Reihe zu

$$\hat{f}(\hat{z}) = c_0 + 2\sum_{k=1}^{\infty} c_k \frac{\hat{z}^k + \hat{z}^{-k}}{2}$$

umzuschreiben.
Da jedoch Aussagen zur Interpolation auf den Bernstein-Ellipsen und nicht auf Kreisringen gesucht werden, nutze für alle $z \in D_\varrho$ die Existenz einer Rechtsinversen 4.5 um $\hat{z} = \varpi^\dagger(z)$ einzusetzen und

$$\hat{f}(\hat{z}) = \hat{f}(\varpi^\dagger(z)) = f(\varpi(\varpi^\dagger(z))) = f(z)$$

zu erhalten. Auf der rechten Seite der Laurent-Reihe treten Summanden auf, die sich auf Tschebyscheff-Polynome zurückführen lassen [37, S. 150].

Lemma 4.7 (Darstellung Tschebyscheff-Polynome)
Für $z \in \mathbb{C}$ mit $\varpi^\dagger(z) = x$ gilt

$$T_n(z) = \frac{x^n + x^{-n}}{2} \qquad \textit{für alle } n \in \mathbb{N}_0.$$

Beweis: Führe den Beweis per Induktion.
I.A. Für $n = 0$ gilt

$$T_0(z) = 1 = \frac{2}{2} = \frac{1+1}{2} = \frac{x^n + x^{-n}}{2}$$

und für $n = 1$ ergibt sich

$$T_1(z) = z = \varpi(\varpi^\dagger(z)) = \varpi(x) = \frac{x + x^{-1}}{2} = \frac{x^n + x^{-n}}{2}.$$

I.V. Sei $n \in \mathbb{N}$ so gegeben, dass die Behauptung gilt.
I.S. Betrachte $n + 1$. Da dann $n + 1 \geq 2$ erfüllt ist, kann T_{n+1} mit der Rekursion für die Tschebyscheff-Polynome (2.1.3) geschrieben werden, es folgt

$$\begin{aligned}
T_{n+1}(z) &= 2zT_n(z) - T_{n-1}(z) = 2\varpi(x)T_n(z) - T_{n-1}(z) \\
&\overset{I.V.}{=} 2\varpi(x)\frac{x^n + x^{-n}}{2} - \frac{x^{n-1} + x^{-(n-1)}}{2} \\
&= (x + x^{-1})\frac{x^n + x^{-n}}{2} + \frac{x^{n-1} - x^{-(n-1)}}{2} \\
&= \frac{x^{n+1} + x^{-(n-1)} + x^{n-1} + x^{-(n+1)}}{2} - \frac{x^{n-1} + x^{-(n-1)}}{2} \\
&= \frac{x^{n+1} + x^{-(n+1)}}{2}.
\end{aligned}$$

□[8]

[iv] Nach dem französischen Mathematiker Augustin-Louis Cauchy benannt.

Mit der Darstellung der Tschebyscheff-Polynome aus dem Lemma 4.7 folgt für alle $z \in D_\varrho$ zusammen mit der Existenz einer Rechtsinversen 4.5, welche $\hat{z} = \varpi^\dagger(z)$ liefert,

$$\tfrac{\hat{z}^k + \hat{z}^{-k}}{2} = T_k(z).$$

Für die Laurent-Reihe gilt somit

$$f(z) = c_0 + 2\sum_{k=1}^{\infty} c_k T_k(z) \qquad \text{für alle } z \in D_\varrho.$$

Ein Approximationspolynom kann über Abschneiden der Laurent-Reihe gewonnen werden. Um den Fehler des Abschneidens abschätzen zu können, sind Schranken für die Koeffizienten und die Tschebyscheff-Polynome nötig.
Für die Koeffizienten ergibt sich für $r \in (\varrho^{-1}, \varrho)$

$$\begin{aligned}
|c_k| &= \left| \tfrac{1}{2i\pi} \int_{|s|=r} \tfrac{\hat{f}(s)}{s^{k+1}} \,\mathrm{d}s \right| \overset{\Delta}{\leq} \tfrac{1}{2\pi} \int_{|s|=r} \tfrac{|\hat{f}(s)|}{|s|^{k+1}} \,\mathrm{d}s \\
&\leq \tfrac{1}{2\pi} \max\left\{ |\hat{f}(s)| \mid s \in \mathbb{C}, |s| = r \right\} \int_{|s|=r} \tfrac{1}{|s|^{k+1}} \,\mathrm{d}s \\
&= \tfrac{1}{2\pi} \max\left\{ |\hat{f}(s)| \mid s \in \mathbb{C}, |s| = r \right\} \int_{|s|=r} \tfrac{1}{r^{k+1}} \,\mathrm{d}s \\
&= \tfrac{1}{2\pi} 2\pi r \tfrac{1}{r^{k+1}} \max\left\{ |\hat{f}(s)| \mid s \in \mathbb{C}, |s| = r \right\} \\
&= \tfrac{1}{r^k} \max\left\{ |\hat{f}(s)| \mid s \in \mathbb{C}, |s| = r \right\}.
\end{aligned}$$

Da $r \in (\varrho^{-1}, \varrho)$ für die Cauchy-Integralformel beliebig war, kann eine Grenzwertbetrachtung gemacht und $\hat{f}$ auf f zurückgeführt werden

$$|c_k| \leq \lim_{r \to \varrho} \left(\tfrac{1}{r^k} \max\left\{ |\hat{f}(s)| \mid s \in \mathbb{C}, |s| = r \right\} \right) \leq \tfrac{\|\hat{f}\|_{\infty, A_\varrho}}{\varrho^k} = \tfrac{\|f\|_{\infty, D_\varrho}}{\varrho^k}. \tag{4.1.7}$$

Für ein $\widehat{\varrho} \in (1, \varrho)$ und $z \in D_{\widehat{\varrho}}$ gilt für $\hat{z} = \varpi^\dagger(z) \in A_{\widehat{\varrho}}$ sowie $\widehat{\varrho}^{-1} < |\hat{z}| < \widehat{\varrho}$. Damit folgt für die Tschebyscheff-Polynome

$$|T_k(z)| = \left| \tfrac{\hat{z}^k + \hat{z}^{-k}}{2} \right| \overset{\Delta}{\leq} \tfrac{|\hat{z}|^k + |\hat{z}^{-1}|^k}{2} \leq \tfrac{\widehat{\varrho}^k + \widehat{\varrho}^k}{2} = \tfrac{2\widehat{\varrho}^k}{2} = \widehat{\varrho}^k. \tag{4.1.8}$$

Wird das Tschebyscheff-Polynom auf dem Referenzintervall $[-1, 1]$ betrachtet, ergibt sich sofort mit der Darstellung über den Kosinus (2.1.4)

$$|T_k(x)| \leq 1 \qquad \text{für alle } x \in [-1, 1]. \tag{4.1.9}$$

Mit diesen Schranken kann eine Fehleraussage zur Approximation holomorpher Funktionen auf Bernstein-Ellipsen und dem Einheitskreis getroffen werden.

Satz 4.8 (Holomorphe Approximation)
Seien $\varrho \in \mathbb{R}_{>1}$, $\widehat{\varrho} \in (1, \varrho)$ und eine holomorphe Funktion $f : D_\varrho \to \mathbb{C}$ gegeben. Für $m \in \mathbb{N}_0$ und das Polynom

$$p = c_0 + 2\sum_{k=1}^{m} c_k T_k$$

ergibt sich ein Approximationsfehler von

$$\|f - p\|_{\infty, D_{\widehat{\varrho}}} \leq \tfrac{2}{\varrho\widehat{\varrho}^{-1} - 1} \left(\tfrac{\widehat{\varrho}}{\varrho}\right)^m \|f\|_{\infty, D_\varrho}.$$

Insbesondere folgt auf dem Referenzintervall $[-1, 1]$

$$\|f - p\|_{\infty,[-1,1]} \leq \tfrac{2}{\varrho - 1} \varrho^{-m} \|f\|_{\infty, D_\varrho}.$$

Beweis: Mit der Laurent-Reihe der Funktion f ergibt sich für den Fehler in $z \in D_{\widehat{\varrho}}$

$$\begin{aligned} |f(z) - p(z)| &= \left| c_0 + 2\sum_{k=1}^{\infty} c_k T_k(z) - c_0 - 2\sum_{k=1}^{m} c_k T_k(z) \right| = \left| 2\sum_{k=m+1}^{\infty} c_k T_k(z) \right| \\ &\overset{\Delta}{\leq} 2\sum_{k=m+1}^{\infty} |c_k| \|T_k\|_{\infty, D_{\widehat{\varrho}}}. \end{aligned}$$

Einsetzen der Schranken für die Tschebyscheff-Polynome (4.1.8) und die Koeffizienten (4.1.7) liefert

$$\begin{aligned} 2\sum_{k=m+1}^{\infty} |c_k| \|T_k\|_{\infty, D_{\widehat{\varrho}}} &\leq 2\sum_{k=m+1}^{\infty} \tfrac{\|f\|_{\infty, D_\varrho}}{\varrho^k} \widehat{\varrho}^k = 2\|f\|_{\infty, D_\varrho} \sum_{k=m+1}^{\infty} \left(\tfrac{\widehat{\varrho}}{\varrho}\right)^k \\ &= 2\|f\|_{\infty, D_\varrho} \left(\tfrac{\widehat{\varrho}}{\varrho}\right)^{m+1} \sum_{k=0}^{\infty} \left(\tfrac{\widehat{\varrho}}{\varrho}\right)^k. \end{aligned}$$

Da nach der Wahl von $\widehat{\varrho}$ grundsätzlich $\widehat{\varrho} < \varrho$ gilt, kann die unendliche Summe mit Hilfe des Grenzwerts der geometrischen Reihe abgeschätzt werden

$$\left(\tfrac{\widehat{\varrho}}{\varrho}\right)^{m+1} \sum_{k=0}^{\infty} \left(\tfrac{\widehat{\varrho}}{\varrho}\right)^k = \left(\tfrac{\widehat{\varrho}}{\varrho}\right)^{m+1} \tfrac{1}{1 - \widehat{\varrho}\varrho^{-1}} = \left(\tfrac{\widehat{\varrho}}{\varrho}\right)^m \tfrac{1}{\varrho\widehat{\varrho}^{-1} - 1}$$

und damit folgt die erste Behauptung.
Wird der Fehler auf dem Referenzintervall betrachtet, folgt für $x \in [-1, 1]$ mit den Schranken (4.1.9) und (4.1.7)

$$\begin{aligned} |f(x) - p(x)| &\leq 2\sum_{k=m+1}^{\infty} |c_k| \|T_k\|_{\infty,[-1,1]} \leq 2\|f\|_{\infty, D_\varrho} \sum_{k=m+1}^{\infty} \left(\tfrac{1}{\varrho}\right)^k \\ &= 2\|f\|_{\infty, D_\varrho} \left(\tfrac{1}{\varrho}\right)^{m+1} \sum_{k=0}^{\infty} \left(\tfrac{1}{\varrho}\right)^k. \end{aligned}$$

Dies kann wegen $\varrho > 1$ erneut mit dem Grenzwert der geometrischen Reihe abgeschätzt werden, so dass

$$\|f - p\|_{\infty,[-1,1]} \leq \|f\|_{\infty,D_\varrho} \varrho^{-m} \tfrac{2}{\varrho - 1}$$

folgt. □[8]

Bemerkung 20 : *Das Maximumprinzip liefert, dass eine auf D_ϱ holomorphe und bis zum Rand ∂D_ϱ stetige Funktion f ihr Maximum auf dem Rand annimmt. Entsprechend wird bei konkreten Bestimmungen der Fehler mit dem Satz 4.8 direkt*

$$\max\left\{|f(z)| \mid z \in \overline{D}_\varrho\right\}$$

berechnet, da die dort betrachteten Funktionen f auf größeren Gebieten holomorph und damit auf ∂D_ϱ stetig sind.

Der letzte Satz zusammen mit der Bernstein-Ungleichung [20, S. 101 Thm. 2.2] für $m \in \mathbb{N}_0$

$$\|q\|_{\infty,D_\varrho} \leq \varrho^m \|q\|_{\infty,[-1,1]} \qquad \text{für alle } q \in \Pi_m \tag{4.1.10}$$

kann zu einer Aussage über den Interpolationsfehler auf Bernstein-Ellipsen kombiniert werden. Auch diese Aussage stammt aus einer unveröffentlichten Arbeit von Herrn Börm [9].

Lemma 4.9 (Interpolationsfehler auf der Bernstein-Ellipse)
Seien $\varrho \in \mathbb{R}_{>1}$ und $\widehat{\varrho} \in (1, \varrho)$ gegeben. Für eine holomorphe Funktion $f : D_\varrho \to \mathbb{C}$ ist der Fehler der Interpolation mit Ordnung $m \in \mathbb{N}_0$ durch

$$\left\|f - \mathfrak{I}_{[-1,1]}[f]\right\|_{\infty,D_{\widehat{\varrho}}} \leq (1 + \Lambda_m) \tfrac{2}{\varrho\widehat{\varrho}^{-1} - 1} \left(\tfrac{\widehat{\varrho}}{\varrho}\right)^m \|f\|_{\infty,D_\varrho}$$

sowie

$$\left\|f - \mathfrak{I}_{[-1,1]}[f]\right\|_{\infty,[-1,1]} \leq (1 + \Lambda_m) \tfrac{2}{\varrho - 1} \varrho^{-m} \|f\|_{\infty,D_\varrho}$$

beschränkt.

Beweis: Zunächst kann das Polynom $p \in \Pi_m$ aus dem Satz 4.8 eingeschoben werden

$$\begin{aligned}
\left\|f - \mathfrak{I}_{[-1,1]}[f]\right\|_{\infty,D_{\widehat{\varrho}}} &= \left\|f - p + p - \mathfrak{I}_{[-1,1]}[f]\right\|_{\infty,D_{\widehat{\varrho}}} \\
&\overset{\Delta}{\leq} \|f - p\|_{\infty,D_{\widehat{\varrho}}} + \left\|p - \mathfrak{I}_{[-1,1]}[f]\right\|_{\infty,D_{\widehat{\varrho}}} \\
&= \|f - p\|_{\infty,D_{\widehat{\varrho}}} + \left\|\mathfrak{I}_{[-1,1]}[f - p]\right\|_{\infty,D_{\widehat{\varrho}}}.
\end{aligned}$$

Der zweite Summand kann mit der Bernstein-Ungleichung (4.1.10) und der Stabilität des Interpolationsoperators (siehe Definition 2.3) gegen

$$\left\|\mathfrak{I}_{[-1,1]}[f - p]\right\|_{\infty,D_{\widehat{\varrho}}} \leq \widehat{\varrho}^m \left\|\mathfrak{I}_{[-1,1]}[f - p]\right\|_{\infty,[-1,1]} \leq \widehat{\varrho}^m \Lambda_m \|f - p\|_{\infty,[-1,1]}$$

abgeschätzt werden. Auf beide verbleibende Normen kann der Satz 4.8 angewendet werden. Es folgt

$$\|f-p\|_{\infty,D_{\widehat{\varrho}}} \leq \tfrac{2}{\varrho\widehat{\varrho}^{-1}-1}\left(\tfrac{\widehat{\varrho}}{\varrho}\right)^m \|f\|_{\infty,D_\varrho}$$

und da $1 < \varrho\widehat{\varrho}^{-1} < \varrho$ erfüllt ist, ergibt sich

$$\widehat{\varrho}^m\Lambda_m\|f-p\|_{\infty,[-1,1]} \leq \Lambda_m \tfrac{2}{\varrho-1}\left(\tfrac{\widehat{\varrho}}{\varrho}\right)^m \|f\|_{\infty,D_\varrho} \leq \Lambda_m \tfrac{2}{\varrho\widehat{\varrho}^{-1}-1}\left(\tfrac{\widehat{\varrho}}{\varrho}\right)^m \|f\|_{\infty,D_\varrho}.$$

Die Summe der beiden Abschätzungen liefert die erste Behauptung.
Auch im zweiten Fall kann das Polynom $p \in \Pi_m$ aus dem Satz 4.8 eingeschoben

$$\begin{aligned}\left\|f-\mathfrak{I}_{[-1,1]}[f]\right\|_{\infty,[-1,1]} &= \left\|f-p+p-\mathfrak{I}_{[-1,1]}[f]\right\|_{\infty,[-1,1]} \\ &\overset{\Delta}{\leq} \|f-p\|_{\infty,[-1,1]} + \left\|p-\mathfrak{I}_{[-1,1]}[f]\right\|_{\infty,[-1,1]} \\ &= \|f-p\|_{\infty,[-1,1]} + \left\|\mathfrak{I}_{[-1,1]}[f-p]\right\|_{\infty,[-1,1]}\end{aligned}$$

und der zweite Summand mit der Stabilitätskonstante des Interpolationsoperators gegen

$$\left\|\mathfrak{I}_{[-1,1]}[f-p]\right\|_{\infty,[-1,1]} \leq \Lambda_m\|f-p\|_{\infty,[-1,1]}$$

abgeschätzt werden. Anschließend nutze die Fehleraussage aus Satz 4.8, um

$$(1+\Lambda_m)\|f-p\|_{\infty,[-1,1]} \leq (1+\Lambda_m)\tfrac{2}{\varrho-1}\varrho^{-m}\|f\|_{\infty,D_\varrho}$$

und damit dann die Behauptung zu erhalten. □

Auch wenn nur auf $[-1,1]$ interpoliert werden soll, liefert dieser Satz die Aussage, dass die Interpolation einer holomorphen Funktion sogar auf einer Bernstein-Ellipse, also einer Umgebung von $[-1,1]$, noch verwendbare Ergebnisse liefert.

4.2 Holomorphe Fortsetzungen der Kernfunktionen

Die Analyse des Tensorinterpolationsansatzes auf dem Gebiet $Q_t \times Q_s$ kann mit Hilfe des Lemmas 2.7 beziehungsweise des Satzes 2.8 auf die Fehlerbetrachtung im Eindimensionalen zurückgeführt werden. Um den Fehler in einer Koordinate abschätzen zu können, wird der Satz 4.8 verwendet. Voraussetzung, um mit diesem Ansatz eine Aussage zum Interpolationsfehler auf $[-1,1]$ zu erhalten, ist es, dass die betrachtete Funktion auf einer Bernstein-Ellipse holomorph ist.
Entsprechend beschäftigt sich dieser Abschnitt damit, holomorphe Fortsetzungen der Kernfunktionen auf Teilgebieten von $\mathbb{C}$ zu finden und auf Formulierungen mit eindimensionalem Definitionsbereich zurückzuführen.

Der ebenen Welle kommt bei der Fehlerbetrachtung eine besondere Stellung zu, denn für einen Block $(t, s, c) \in \mathcal{T}_{\mathcal{I}\times\mathcal{I}}$ gilt für alle $x \in Q_t$ und $y \in Q_s$ mit der Euler'schen Formel[v]

$$\begin{aligned}\left|e^{\mathrm{i}\kappa\langle x-y,c\rangle_2}\right|^2 &= |\cos(\kappa\langle x-y,c\rangle_2) + \mathrm{i}\sin(\kappa\langle x-y,c\rangle_2)|^2\\ &= \cos^2(\kappa\langle x-y,c\rangle_2) + \sin^2(\kappa\langle x-y,c\rangle_2).\end{aligned}$$

Damit folgt

$$\left|e^{\mathrm{i}\kappa\langle x-y,c\rangle_2}\right| = 1, \tag{4.2.1}$$

was die Möglichkeit eröffnet, die ebene Welle trotz ihres oszillierenden Charakters im Betrag in einigen Betrachtungen einfach wegzulassen oder hinzuzufügen.
Um mit eindimensionalen Definitionsbereichen arbeiten zu können, wird die Idee der bijektiven Abbildung Φ (2.1.6), welche es erlaubt, vom Einheitsintervall aus zu agieren, übertragen und angepasst. Da die Kernfunktionen von der Differenz von x, y abhängig sind, behandele die Variablen x, y als eine Einheit und führe zwei dreidimensionale Vektoren $\mathfrak{m}, \mathfrak{l} \in \mathbb{R}^3$ ein, die in einer Komponente den Intervallmittelpunkt beziehungsweise die Intervalllänge des zu betrachtenden überdeckenden Quaders Q_t oder Q_s enthalten. Für ein festes $i \in \underline{6}$ ist der Vektor $\mathfrak{m}$ zum Intervallmittelpunkt eintragsweise definiert durch

$$\mathfrak{m}_j := \begin{cases} \frac{b_{t,j}+a_{t,j}}{2} - y_j & \text{falls } i = j,\\ x_j - \frac{b_{s,j}+a_{s,j}}{2} & \text{falls } i = j+3, \qquad \text{für alle } j \in \underline{3},\\ x_j - y_j & \text{sonst}\end{cases}$$

während der Vektor $\mathfrak{l}$ zur halben Intervalllänge eintragsweise durch

$$\mathfrak{l}_j := \begin{cases} -\frac{b_{t,j}-a_{t,j}}{2} & \text{falls } i = j,\\ \frac{b_{s,j}-a_{s,j}}{2} & \text{falls } i = j+3, \qquad \text{für alle } j \in \underline{3}\\ 0 & \text{sonst}\end{cases}$$

gegeben ist. Auf Grund der Konstruktion erfüllen die Vektoren $\mathfrak{l}, \mathfrak{m}$ (vgl. [3, Lem. 3.4])

$$\|\mathfrak{l}\|_2 \leq \tfrac{\max\{\operatorname{diam}(Q_t),\operatorname{diam}(Q_s)\}}{2} \tag{4.2.2a}$$

$$\mathfrak{m} - \tau\mathfrak{l} \in \{\widetilde{x} - \widetilde{y} \mid \widetilde{x} \in Q_t, \widetilde{y} \in Q_s\} = Q_t - Q_s \qquad \text{für alle } \tau \in [-1, 1]. \tag{4.2.2b}$$

Für ein $i \in \underline{6}$ mit wie oben definierten $\mathfrak{m}, \mathfrak{l}$ kann die Analyse der Tensorinterpolation auf $Q_t \times Q_s$ wie in Lemma 2.7 auf die Betrachtung des eindimensionalen Interpolationsoperators $\mathfrak{I}_{[-1,1]}$ zurückgeführt werden.
Um die holomorphen Erweiterungen der beteiligten Kernfunktionen zu bilden, ist es zudem

[v] Der schweizerische Mathematiker und Physiker Leonhard Euler entdeckte den Zusammenhang der trigonometrischen Funktionen und der komplexen Exponentialfunktion.

erforderlich, dass sowohl das Skalarprodukt als auch die Norm adäquat im Komplexen fortgesetzt werden. Das Übertragen der Exponentialfunktion ist kein Problem, da sie mit Hilfe ihrer Reihendarstellung direkt fortgesetzt werden kann. Wenn durch $\langle \cdot, \cdot \rangle_E$ eine Fortsetzung des Skalarprodukts und durch $\| \cdot \|_E$ eine Fortsetzung der Norm im Komplexen gegeben sind, dann kann die richtungsabhängige Variante der Kernfunktion für die Interpolation entlang einer Koordinate $i \in \underline{6}_1$ im Fall des Einfachschichtoperators mit

$$ {}^i g_{ec}(\tau) = \frac{e^{\mathrm{i}\kappa\left(\|\mathfrak{m}-\tau\mathfrak{l}\|_E - \langle \mathfrak{m}-\tau\mathfrak{l}, c\rangle_E\right)}}{4\pi\|\mathfrak{m}-\tau\mathfrak{l}\|_E} \qquad \text{für alle } \tau \in [-1,1] \tag{4.2.3} $$

und im Fall des Doppelschichtoperators für die Ableitung in y_j mit $j \in \underline{3}_1$ und $j' = j+3$ mit

$$ {}^i g_{dc,j'}(\tau) := {}^i g_{ec}(\tau) \left(\frac{\langle \mathfrak{m}-\tau\mathfrak{l}, e_j\rangle_E}{\|\mathfrak{m}-\tau\mathfrak{l}\|_E} \left(\frac{1}{\|\mathfrak{m}-\tau\mathfrak{l}\|_E} - \mathrm{i}\kappa \right) + \mathrm{i}\kappa \langle c, e_j \rangle_2 \right) \qquad \text{für alle } \tau \in [-1,1] \tag{4.2.4} $$

geschrieben werden, wobei $\mathfrak{m}, \mathfrak{l}$ jeweils (4.2.2) erfüllen.

Bleiben die Fragen, wie die komplexen Fortsetzungen $\langle \cdot, \cdot \rangle_E$ und $\| \cdot \|_E$ zu definieren und wie die Definitionsbereiche zu wählen sind, so dass die Kernfunktionen dann auf den gewählten Definitionsbereichen holomorph sind [3, Kap. 3.2].
Beginne mit dem Skalarprodukt und definiere die Fortsetzung des reellen euklidischen Skalarprodukts durch

$$ \langle x, y \rangle_E := \sum_{i=1}^{d} x_i y_i \qquad \text{für alle } x, y \in \mathbb{C}^d. $$

Durch $\langle \cdot, \cdot \rangle_E$ ist jedoch kein Skalarprodukt auf $\mathbb{C}^d$ gegeben, denn $\langle x, x \rangle_E$ ist nicht für alle $x \in \mathbb{C}^d \setminus \mathbb{R}^d$ positiv.

Lemma 4.10 (Eigenschaft fortgesetztes Skalarprodukt)
Seien $\mathfrak{m}, \mathfrak{l} \in \mathbb{R}^3$ mit $\mathfrak{l} \neq 0$ gegeben, es gilt:

$$ \langle \mathfrak{m} - z\mathfrak{l}, \mathfrak{m} - z\mathfrak{l} \rangle_E = \|\mathfrak{l}\|_2^2 (\omega - z)(\overline{\omega} - z) \qquad \textit{für alle } z \in \mathbb{C} $$

mit $\omega = \omega_r + \mathrm{i}\omega_i$ und

$$ \omega_r := \frac{\langle \mathfrak{m}, \mathfrak{l} \rangle_2}{\|\mathfrak{l}\|_2^2} \qquad \textit{sowie} \qquad \omega_i := \sqrt{\frac{\|\mathfrak{m}\|_2^2}{\|\mathfrak{l}\|_2^2} - \omega_r^2}. \tag{4.2.5} $$

Beweis: Sei ein $z \in \mathbb{C}$ gegeben, dann gilt für das erweiterte Skalarprodukt mit $\mathfrak{m}, \mathfrak{l} \in \mathbb{R}^3$

$$ \begin{aligned} \langle \mathfrak{m} - z\mathfrak{l}, \mathfrak{m} - z\mathfrak{l} \rangle_E &= \sum_{i=1}^{3} (\mathfrak{m}_i - z\mathfrak{l}_i)^2 = \sum_{i=1}^{3} \mathfrak{m}_i^2 - 2z \sum_{i=1}^{3} \mathfrak{m}_i \mathfrak{l}_i + z^2 \sum_{i=1}^{3} \mathfrak{l}_i^2 \\ &= \|\mathfrak{m}\|_2^2 - 2z\langle \mathfrak{m}, \mathfrak{l} \rangle_2 + z^2 \|\mathfrak{l}\|_2^2. \end{aligned} $$

Für ω mit ω_r, ω_i nach (4.2.5) ergibt sich $|\omega| = \frac{\|\mathfrak{m}\|_2}{\|\mathfrak{l}\|_2}$ sowie für ω_r

$$2\omega_r = \omega_r + \mathrm{i}\omega_i - \mathrm{i}\omega_i + \omega_r = \omega + \overline{\omega}.$$

All dies zusammen kann genutzt werden, um

$$\begin{aligned}\|\mathfrak{m}\|_2^2 - 2z\langle\mathfrak{m},\mathfrak{l}\rangle_2 + z^2\|\mathfrak{l}\|_2^2 &= |\omega|^2\|\mathfrak{l}\|_2^2 - 2z\|\mathfrak{l}\|_2^2\omega_r + z^2\|\mathfrak{l}\|_2^2 = \|\mathfrak{l}\|_2^2(|\omega|^2 - 2z\omega_r + z^2)\\ &= \|\mathfrak{l}\|_2^2(|\omega|^2 - z(\omega+\overline{\omega}) + z^2) = \|\mathfrak{l}\|_2^2(\omega - z)(\overline{\omega} - z)\end{aligned} \tag{4.2.6}$$

zu erhalten. □[3]

Die Fortsetzung der Norm bedarf einer Fortsetzung der Wurzel, was sich deutlich komplizierter gestaltet, da die Wurzel nicht einheitlich auf gesamt $\mathbb{C}$ fortsetzbar ist. Für den hier betrachteten Zweck reicht eine Erweiterung auf den Hauptzweig der komplexen Wurzel aus [3, Kap. 3.2], definiere die Fortsetzung der Wurzel mit

$$\sqrt{z} := \sqrt{|z|}\tfrac{z+|z|}{|z+|z||} \qquad \text{für alle } z \in \mathbb{C}\setminus\mathbb{R}_{\leq 0}. \tag{4.2.7}$$

Die rechte Seite ist wohldefiniert, da für den Term im Nenner

$$\begin{aligned}|z+|z|| = 0 &\iff |z+|z||^2 = 0 \iff (z+|z|)\overline{(z+|z|)} = 0\\ &\iff |z|(2|z| + z + \overline{z}) = 0\end{aligned}$$

gilt und dies nur dann erfüllt sein kann, wenn $z = 0$ oder

$$2|z| + z + \overline{z} = 0 \iff 2|z| = -(z+\overline{z}) = -2\Re(z)$$

gilt, also wenn $z \in \mathbb{R}_{\leq 0}$. Außerdem erfüllt die Definition ihren Zweck als Fortsetzung der Wurzel, denn es gilt

$$\begin{aligned}(\sqrt{z})^2 = \left(\sqrt{|z|}\tfrac{z+|z|}{|z+|z||}\right)^2 &= |z|\tfrac{(z+|z|)^2}{(z+|z|)(z+|z|)} = |z|\tfrac{z^2+2z|z|+z\overline{z}}{z\overline{z}+z|z|+\overline{z}|z|+|z|^2}\\ &= |z|\tfrac{z(z+2|z|+\overline{z})}{|z|(|z|+z+\overline{z}+|z|)} = \tfrac{z(z+2|z|+\overline{z})}{z+\overline{z}+2|z|} = z.\end{aligned}$$

Da die Fortsetzung der Norm einer komplexen Zahl z immer mit gegebenen $\mathfrak{m}, \mathfrak{l}$ in der Form $\|\mathfrak{m} - z\mathfrak{l}\|_E$ auftritt, ist es sinnvoll, den Definitionsbereich in Abhängigkeit von $\mathfrak{m}, \mathfrak{l}$ zu bestimmen und die Norm dann mit

$$\|\mathfrak{m} - z\mathfrak{l}\|_E := \sqrt{\langle\mathfrak{m} - z\mathfrak{l}, \mathfrak{m} - z\mathfrak{l}\rangle_E}$$

zu definieren. Als Definitionsbereich der fortgesetzten Norm bietet sich für $\mathfrak{m}, \mathfrak{l} \in \mathbb{R}^3$ mit $\mathfrak{l} \neq 0$ und ω wie in (4.2.5) die Menge

$$\Upsilon_{\mathfrak{m}\mathfrak{l}} := \mathbb{C}\setminus\{\omega_r + \mathrm{i}b \mid b \in \mathbb{R}, |b| \geq \omega_i\}$$

an [3, Lem 3.5]. Die Wahl des Definitionsbereichs $\Upsilon_{\mathfrak{m}\mathfrak{l}}$ resultiert aus der nachfolgenden Überlegung.
Starte die Betrachtung dazu bei der mittleren Formulierung von (4.2.6) und zerlege z mit passenden $a, b \in \mathbb{R}$ in der Form $z = a + \mathrm{i}b$

$$\begin{aligned}\|\mathfrak{l}\|_2^2(|\omega|^2 - 2z\omega_r + z^2) &= \|\mathfrak{l}\|_2^2\left(|\omega|^2 - 2(a+\mathrm{i}b)\omega_r + (a+\mathrm{i}b)^2\right)\\ &= \|\mathfrak{l}\|_2^2\left(|\omega|^2 + a^2 - 2a\omega_r - b^2 - 2\mathrm{i}b(\omega_r - a)\right).\end{aligned}$$

Damit die fortgesetzte Wurzelfunktion Anwendung finden darf, müssen nach (4.2.7) einige $z \in \mathbb{C}$ ausgeschlossen werden. Um diese z zu ermitteln, reicht es, den zweiten Faktor näher zu betrachten und Bedingungen zu finden, bei denen der zweite Faktor in $\mathbb{R}_{\leq 0}$ liegt. Das Verschwinden des Imaginärteils und das Vorhandensein eines nicht positiven Realteils erfordern, dass $b = 0$ oder $a = \omega_r$ ist sowie dass zusätzlich

$$|\omega|^2 + a^2 - 2a\omega_r - b^2 \leq 0 \qquad \Longleftrightarrow \qquad |\omega|^2 + a^2 - 2a\omega_r \leq b^2$$

gilt.
Beginne mit dem Fall $b \neq 0$ und $a = \omega_r$, welcher zur Ungleichung

$$b^2 \geq |\omega|^2 + a^2 - 2a\omega_r = |\omega|^2 - \omega_r^2 = \omega_i^2$$

führt. Wird der Fall $b = 0$ betrachtet, ergibt sich durch Ausschreiben von $|\omega|^2$

$$0 \geq |\omega|^2 + a^2 - 2a\omega_r = \omega_r^2 + \omega_i^2 + a^2 - 2a\omega_r = (a - \omega_r)^2 + \omega_i^2,$$

was dann

$$-\omega_i^2 \geq (a - \omega_r)^2$$

liefert. Dies ist nur dann erfüllt, wenn beide Seiten null sind, womit erneut $a = \omega_r$ gelten muss. Entsprechend reicht die Forderung

$$|b| \geq \omega_i$$

in $\Upsilon_{\mathfrak{m}\mathfrak{l}}$ aus, um den Definitionsbereich einzuschränken, denn $\langle \mathfrak{m} - z\mathfrak{l}, \mathfrak{m} - z\mathfrak{l}\rangle_E \in \mathbb{R}_{\leq 0}$ impliziert, dass $z \notin \Upsilon_{\mathfrak{m}\mathfrak{l}}$ gilt.

Um die Notation für die hier vorgenommene Betrachtung zu vereinfachen, führe spezielle Funktionen für die fortgesetzte Norm und den Exponentenanteil ein.

Definition 4.11 (Fortgesetzte Norm und Exponentenfunktion)
Seien $\mathfrak{m}, \mathfrak{l} \in \mathbb{R}^3$ mit $\mathfrak{l} \neq 0$ und ω wie in (4.2.5) *gegeben. Definiere die fortgesetzte Norm*

$$f_{\|\cdot\|} : \Upsilon_{\mathfrak{m}\mathfrak{l}} \to \mathbb{C}_+ = \{z \in \mathbb{C} \mid \Re(z) > 0\}, \qquad z \mapsto \sqrt{\langle \mathfrak{m} - z\mathfrak{l}, \mathfrak{m} - z\mathfrak{l}\rangle_E}$$

sowie die Exponentenfunktion mit einem Richtungsvektor $c \in \mathbb{R}^3$, $\|c\|_2 = 1$

$$f_e : \Upsilon_{\mathfrak{m}\mathfrak{l}} \to \mathbb{C}, \qquad z \mapsto f_{\|\cdot\|}(z) - \langle \mathfrak{m} - z\mathfrak{l}, c\rangle_E.$$

Die Holomorphie der fortgesetzten Normfunktion folgt direkt aus den bisherigen Betrachtungen.

Lemma 4.12 (Holomorphie der Normfunktion)
Seien $\mathfrak{m}, \mathfrak{l} \in \mathbb{R}^3$ *mit* $\mathfrak{l} \neq 0$ *gegeben, dann ist die Funktion* $f_{\|\cdot\|}$ *aus Definition 4.11 holomorph auf* $\Upsilon_{\mathfrak{ml}}$*, des Weiteren gilt*

$$f_{\|\cdot\|}(z) = \|\mathfrak{l}\|_2 \sqrt{(\omega - z)(\overline{\omega} - z)} \qquad \text{für alle } z \in \Upsilon_{\mathfrak{ml}},$$

wobei ω *wie in* (4.2.5) *definiert sei.*

Beweis: Die Eigenschaft holomorph zu sein, folgt für $f_{\|\cdot\|}$ direkt daraus, dass die Funktion

$$z \mapsto \langle \mathfrak{m} - z\mathfrak{l}, \mathfrak{m} - z\mathfrak{l} \rangle_E = \|\mathfrak{l}\|_2^2 (\omega - z)(\overline{\omega} - z) \qquad \text{für alle } z \in \mathbb{C}$$

als Polynom holomorph ist. Der Definitionsbereich von $f_{\|\cdot\|}$ wurde so gewählt, dass die komplexe Wurzel, welche auf ihrem Hauptzweig holomorph ist, genutzt wird, und damit die Hintereinanderausführung ebenfalls eine holomorphe Funktion liefert.
Der zweite Teil der Behauptung folgt direkt aus Lemma 4.10 durch Wurzelziehen. □[3]

Die Grenzen der Menge $\Upsilon_{\mathfrak{ml}}$ sind durch ω und $\overline{\omega}$ gegeben, die Menge selbst ist in Abbildung 4.2 für ein $\omega \neq 0$ in blau angedeutet. Bei einem $\omega \notin [-1, 1]$ ist das Intervall $[-1, 1]$ in der Menge $\Upsilon_{\mathfrak{ml}}$ enthalten und damit die hier definierte holomorphe Fortsetzung der Norm für die geplante Analyse verwendbar.

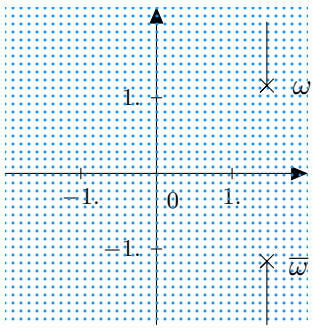
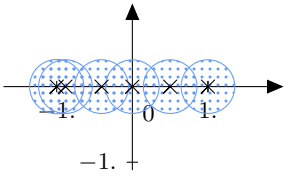


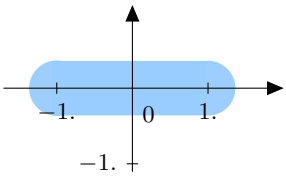
Abbildung 4.2: Menge, auf der die erweiterte Norm holomorph ist

Mit den bisherigen Erkenntnissen kann die Funktion ${}^i g_{ec}$ holomorph erweitert werden, womit die ersten Grundlagen zur Fehleranalyse des Einfachschichtoperators gelegt sind. Im Fall des Doppelschichtoperators kann die Holomorphie auf die Holomorphie der Funktion ${}^i g_{ec}$ sowie der Norm $\|\mathfrak{m} - \tau\mathfrak{l}\|_E$ zurückgeführt werden. Da sowohl Produkte als auch Summen holomorpher Funktionen holomorph sind, ist ${}^i g_{dc,j}$ als Komposition ebenfalls holomorph. Folglich existiert auch eine holomorphe Erweiterung der Funktion ${}^i g_{dc,j}$.

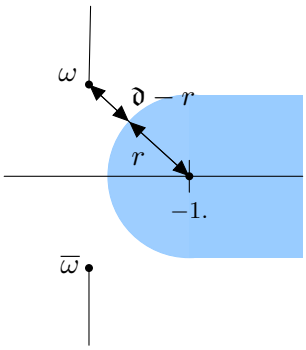
Die Norm tritt in den Kernfunktionen unter anderem im Nenner auf, entsprechend ist eine Abschätzung nach unten von Interesse. Dazu wird ein Teil der Menge $\Upsilon_{\mathfrak{ml}}$ definiert, in dem eine untere Schranke angegeben werden kann. Dies geschieht mit Hilfe von Kugeln, die durch Überlagerung die gesuchte Menge bilden. Das benötigte Intervall $[-1, 1]$ kann um einen Radius, der kleiner als die Distanz zu den Grenzen des Definitionsbereichs der fortgesetzten Wurzel ist, ausgedehnt werden.



(a) Kreise $\Upsilon_r(\tau)$ auf dem Intervall



(b) Teilmenge Υ_r für die untere Schranke



(c) Parameter der Teilmenge

Abbildung 4.3: Betrachtete Teilmenge des Definitionsbereichs

Lemma 4.13 (Normabschätzung)
Seien $\mathfrak{m}, \mathfrak{l} \in \mathbb{R}^3$ *mit* $\mathfrak{l} \neq 0$ *und* $f_{\|\cdot\|}$ *wie in Definition 4.11 gegeben. Definiere zusätzlich Abstände und Kreise für alle* $\tau \in \mathbb{R}$, $r \in \mathbb{R}_{\geq 0}$

$$\mathfrak{d}_\tau := \tfrac{\|\mathfrak{m}-\tau\mathfrak{l}\|_2}{\|\mathfrak{l}\|_2}, \qquad \mathfrak{d} := \min\{\mathfrak{d}_\tau \mid \tau \in [-1,1]\},$$
$$\Upsilon_r(\tau) := \{z \in \mathbb{C} \mid |z-\tau| \leq r\}, \qquad \Upsilon_r := \bigcup_{\tau\in[-1,1]} \Upsilon_r(\tau),$$

dann gilt

$$\Upsilon_r(\tau) \subset \Upsilon_{\mathfrak{ml}} \quad \textit{für alle } r \in [0, \mathfrak{d}_\tau), \qquad \Upsilon_r \subset \Upsilon_{\mathfrak{ml}} \quad \textit{für alle } r \in [0, \mathfrak{d})$$

und die holomorphe Fortsetzung der Norm kann mit

$$\begin{aligned}\left|f_{\|\cdot\|}(z)\right| &\geq \|\mathfrak{l}\|_2(\mathfrak{d}_\tau - r) && \text{für alle } \tau \in \mathbb{R}, r \in [0, \mathfrak{d}_\tau), z \in \Upsilon_r(\tau),\\ \left|f_{\|\cdot\|}(z)\right| &\geq \|\mathfrak{l}\|_2(\mathfrak{d} - r) && \text{für alle } r \in [0, \mathfrak{d}), z \in \Upsilon_r\end{aligned}$$

abgeschätzt werden.

Beweis: Sei ein $\tau \in \mathbb{R}$ gegeben, dann kann der zugehörige Abstand $\mathfrak{d}_\tau$ leicht mit

$$\begin{aligned}\mathfrak{d}_\tau^2 = \tfrac{\|\mathfrak{m}-\tau\mathfrak{l}\|_2^2}{\|\mathfrak{l}\|_2^2} &\overset{4.11}{=} \tfrac{(f_{\|\cdot\|}(\tau))^2}{\|\mathfrak{l}\|_2^2} \overset{4.10}{=} \tfrac{\|\mathfrak{l}\|_2^2(\omega-\tau)(\overline{\omega}-\tau)}{\|\mathfrak{l}\|_2^2} = (\omega-\tau)\overline{(\omega-\tau)}\\ = |\omega-\tau|^2 \qquad &\Longleftrightarrow \qquad \mathfrak{d}_\tau = |\omega-\tau|\end{aligned}$$

bestimmt werden. Der Abstand ermöglicht einen simplen Beweis der Mengeninklusion $\Upsilon_r(\tau) \subset \Upsilon_{\mathfrak{m}\mathfrak{l}}$ per Kontraposition. Entsprechend zeige, dass $\omega_r + \mathrm{i}b \notin \Upsilon_r(\tau)$ für alle $b \in \mathbb{R}$, die $|b| \geq \omega_i$ erfüllen.
Seien $\mathfrak{d}_\tau > 0$, $r \in [0, \mathfrak{d}_\tau)$ und $z = \omega_r + \mathrm{i}b$ mit $|b| \geq \omega_i$ gegeben, es gilt

$$|z-\tau|^2 = (\omega_r-\tau)^2 + b^2 \geq (\omega_r-\tau)^2 + \omega_i^2 = |\omega-\tau|^2 = \mathfrak{d}_\tau^2 > r^2,$$

somit folgt $z \notin \Upsilon_r(\tau)$, was die erste zu zeigende Mengeninklusion impliziert. Da dies für alle $\tau \in [-1,1]$ gilt, folgt direkt, dass dann auch $\Upsilon_r \subset \Upsilon_{\mathfrak{m}\mathfrak{l}}$ erfüllt ist.
Betrachte die erste untere Schranke für die Funktion $f_{\|\cdot\|}$, seien dazu $r \in [0, \mathfrak{d}_\tau)$ und $z \in \Upsilon_r(\tau)$ gegeben, dann gilt mit $|\omega-\tau| = |\overline{\omega}-\tau|$

$$\begin{aligned}|\omega - z| = |\omega-\tau+\tau-z| &\overset{\nabla}{\geq} \big||\omega-\tau| - |\tau-z|\big| \geq |\omega-\tau| - |z-\tau|\\ &= \mathfrak{d}_\tau - |z-\tau| \geq \mathfrak{d}_\tau - r > 0\\ |\overline{\omega} - z| = |\overline{\omega}-\tau+\tau-z| &\overset{\nabla}{\geq} |\overline{\omega}-\tau| - |z-\tau| \geq \mathfrak{d}_\tau - r > 0.\end{aligned}$$

Dies liefert für $z \in \Upsilon_r(\tau)$ insgesamt die erste Abschätzung für die fortgesetzte Norm

$$\left|f_{\|\cdot\|}(z)\right| = \|\mathfrak{l}\|_2\sqrt{|\omega-z||\overline{\omega}-z|} \geq \|\mathfrak{l}\|_2(\mathfrak{d}_\tau - r).$$

Für die zweite Abschätzung seien $r \in [0, \mathfrak{d})$ sowie $z \in \Upsilon_r$ gegeben, dann existiert ein $\tau \in [-1,1]$ mit $z \in \Upsilon_r(\tau)$, weiterhin gilt $\mathfrak{d} = \min\{\mathfrak{d}_\tau \mid \tau \in [-1,1]\} \leq \mathfrak{d}_\tau$. Damit folgt die zweite Abschätzung

$$\left|f_{\|\cdot\|}(z)\right| \geq \|\mathfrak{l}\|_2(\mathfrak{d}_\tau - r) \geq \|\mathfrak{l}\|_2(\mathfrak{d} - r).$$

□[3]

Die Parameter $\mathfrak{m}, \mathfrak{l}$ der Interpolation werden durch Größe, Position und Lage der beiden betrachteten Quader Q_t, Q_s sowie die Wahl der Koordinate $i \in \underline{6}$, entlang welcher die Interpolation untersucht werden soll, festgelegt. Entsprechend ist auch $\mathfrak{d}$ ebenso wie ω kein frei wählbarer Parameter, sondern durch oben genannte Einflüsse festgelegt.

Bemerkung 21 (Zusammenhang zu den Zulässigkeitsbedingungen): *Der Abstand* $\mathfrak{d}$ *kann mit den Zulässigkeitsbedingungen beschränkt werden. Zu* $\mathfrak{d}$ *existiert ein* $\tau \in [-1,1]$*, so dass mit Lemma 4.13 und mit* (4.2.2a) *eine untere Schranke für* $\mathfrak{d}$ *durch*

$$\mathfrak{d} = \mathfrak{d}_\tau = \frac{\|\mathfrak{m}-\tau\mathfrak{l}\|_2}{\|\mathfrak{l}\|_2} \geq \frac{\operatorname{dist}(Q_t,Q_s)}{\|\mathfrak{l}\|_2} \geq \frac{2\operatorname{dist}(Q_t,Q_s)}{\max\{\operatorname{diam}(Q_t),\operatorname{diam}(Q_s)\}}$$

gegeben ist, was mit den Zulässigkeitsbedingungen (2.3.2c)*,* (2.3.2b) *zu*

$$\mathfrak{d} \geq \frac{2}{\eta_2}, \qquad\qquad \mathfrak{d} \geq \frac{4\kappa\|\mathfrak{l}\|_2}{\eta_2}$$

führt.

Die Größe r legt den Radius der Kugeln fest, die überlagert werden (siehe Abbildung 4.3) und auf denen die Abschätzung für die fortgesetzte Norm gilt. Für den Satz zum Approximationsfehler holomorpher Funktionen 4.8 werden jedoch Abschätzungen auf Bernstein-Ellipsen (siehe Def. 4.1) benötigt. Es muss also sichergestellt sein, dass ein Radius r so gefunden werden kann, dass eine dazugehörige Bernstein-Ellipse im Holomorphiegebiet der Funktion liegt und die Abschätzung der fortgesetzten Norm darauf gilt. Der Aufbau des folgenden Beweises ist angelehnt an [4, Lem 4.77].

Lemma 4.14 (Inklusion)
Seien $\mathfrak{m},\mathfrak{l} \in \mathbb{R}^3$ *mit* $\mathfrak{l} \neq 0$ *und* $r \in (0,\mathfrak{d})$ *gegeben. Setze die Halbachsensumme der Bernstein-Ellipse mit* $\varrho = \sqrt{r^2+1}+r$*, dann gilt*

$$\overline{D}_\varrho \subset \Upsilon_{\mathfrak{m}\mathfrak{l}}$$

Beweis: Vorweg werden Beobachtungen zu r und ϱ gemacht, die im Folgenden von Nutzen sein werden.
Starte damit, dass auch ϱ^{-1} mit Hilfe von r ausgedrückt werden kann

$$\sqrt{r^2+1}-r = \frac{\sqrt{r^2+1}-r}{1+r^2-r^2} = \frac{\sqrt{r^2+1}-r}{(\sqrt{r^2+1}+r)(\sqrt{r^2+1}-r)} = \varrho^{-1}.$$

Entsprechend folgt direkt

$$\sqrt{r^2+1} = \frac{2\sqrt{r^2+1}}{2} = \frac{\varrho+\varrho^{-1}}{2} = a_\varrho$$

und

$$r = \frac{2r}{2} = \frac{\varrho-\varrho^{-1}}{2} = b_\varrho.$$

Für ein $z \in \overline{D}_\varrho$ und $a,b \in \mathbb{R}$ mit $z = a+\mathrm{i}b$ gilt damit beim Betrachten der Charakterisierung der Bernstein-Ellipse über die Halbachsen

$$1 \geq \left(\frac{a}{a_\varrho}\right)^2 + \left(\frac{b}{b_\varrho}\right)^2 = \left(\frac{2a}{\varrho+\varrho^{-1}}\right)^2 + \left(\frac{2b}{\varrho-\varrho^{-1}}\right)^2 = \frac{a^2}{r^2+1} + \frac{b^2}{r^2}.$$

Dies kann noch nach a umgestellt werden

$$a^2 \leq (r^2+1)\left(1-\tfrac{b^2}{r^2}\right),$$

was zu

$$a^2 \leq r^2+1-b^2-\tfrac{b^2}{r^2} \leq r^2+1-b^2 \tag{4.2.8}$$

führt. Um zu zeigen, dass das untersuchte z dann auch in $\Upsilon_{\mathfrak{ml}}$ liegt, werden Fallunterscheidungen für a benötigt.
Zu erst sei $a<-1$. Betrachte $(a+1)^2$ und verwende, dass $a<-1$ gilt, um mit 4.2.8

$$(a+1)^2 = a^2+2a+1 \leq r^2+1-b^2+2a+1 < r^2-b^2$$

zu erhalten. Setze $\tau=-1$, damit folgt dann

$$|z-\tau|^2 = |z+1|^2 = (a+1)^2+b^2 < r^2.$$

Entsprechend gilt $|z-\tau|<r$, womit $z\in\Upsilon_r(-1)$ folgt und schließlich mit Lemma 4.13 $\Upsilon_r(-1)\subset\Upsilon_{\mathfrak{ml}}$ und somit $z\in\Upsilon_{\mathfrak{ml}}$.
Falls $a\in[-1,1]$ ist, setze $\tau=a$. Es folgt sofort $|z-\tau|=|b|$ und aus der Charakterisierung der Bernstein-Ellipse ergibt sich $1\geq b^2r^{-2}$ und damit $|b|\leq r$, so dass $z\in\Upsilon_r(\tau)$ folgt.
Zuletzt sei $a>1$, dann setze $\tau=1$. Erneut nutze (4.2.8), um

$$(a-1)^2 = a^2-2a+1 \leq r^2+1-b^2-2a+1 < r^2-b^2$$

zu erhalten. Damit ergibt sich dann

$$|z-\tau|^2 = |z-1|^2 = (a-1)^2+b^2 < r^2,$$

so dass $z\in\Upsilon_r(1)$ folgt. Da für alle $z\in\overline{D}_\varrho$ ein $\tau\in[-1,1]$ gefunden werden kann, so dass $z\in\Upsilon_r(\tau)$ gilt, folgt $\overline{D}_\varrho\subset\Upsilon_{\mathfrak{ml}}$. □

Für die Norm im Nenner der Kernfunktionen ist damit eine zur Interpolation passende Bernstein-Ellipse gefunden, auf der die fortgesetzte Norm holomorph ist und Abschätzungen vorhanden sind. Auch die Holomorphie des modifizierten Arguments der Exponentialfunktion im Zähler ist gezeigt, so dass es nur noch gilt, geeignete Abschätzungen für diese Funktion zu finden. Für die Abschätzung des Exponentialterms findet die Taylor-Entwicklung Anwendung, weshalb zunächst einige nützliche Aussagen zu Ableitungen und Restgliedern gemacht werden sollen.

Lemma 4.15 (Integral)
Seien $\mathfrak{d}>0$ und $r\in[0,\mathfrak{d})$ gegeben, dann gilt

$$\int_0^1 \tfrac{1-s}{(\mathfrak{d}-rs)^3}\,\mathrm{d}s = \tfrac{1}{2\mathfrak{d}^2(\mathfrak{d}-r)}.$$

Beweis: Die Fälle $r = 0$ und $r > 0$ können getrennt behandelt werden, wobei der Fall $r = 0$ besonders leicht ist, denn es folgt sofort

$$\int_0^1 \tfrac{1-s}{\mathfrak{d}^3}\,\mathrm{d}s = \mathfrak{d}^{-3}\int_0^1 (1-s)\,\mathrm{d}s = \tfrac{1}{2\mathfrak{d}^3} = \tfrac{1}{2\mathfrak{d}^2(\mathfrak{d}-r)}.$$

Im zweiten Fall liegt die Herausforderung darin, eine Stammfunktion zu finden. Ein möglicher Kandidat ist durch

$$h : [0,1] \to \mathbb{R}, \qquad\qquad s \mapsto \tfrac{1+\mathfrak{d}r^{-1}-2s}{2r(\mathfrak{d}-rs)^2}$$

gegeben, denn mit der Quotientenregel gilt

$$\begin{aligned} h'(s) &= \tfrac{-2(2r(\mathfrak{d}-rs)^2)-\big(1+\mathfrak{d}r^{-1}-2s\big)2r(-2r)(\mathfrak{d}-rs)}{(2r)^2(\mathfrak{d}-rs)^4} \\ &= \tfrac{-2\mathfrak{d}-2rs+2r+2\mathfrak{d}}{2r(\mathfrak{d}-rs)^3} = \tfrac{1-s}{(\mathfrak{d}-rs)^3}. \end{aligned}$$

Mit dem Hauptsatz der Integralrechnung folgt die Behauptung

$$\begin{aligned} \int_0^1 \tfrac{1-s}{(\mathfrak{d}-rs)^3}\,\mathrm{d}s &= \tfrac{1+\mathfrak{d}r^{-1}-2}{2r(\mathfrak{d}-r)^2} - \tfrac{1+\mathfrak{d}r^{-1}}{2r(\mathfrak{d})^2} \\ &= \tfrac{1}{2}\left(\tfrac{(\mathfrak{d}r^{-1}-1)\mathfrak{d}^2-(1+\mathfrak{d}r^{-1})(\mathfrak{d}-r)^2}{r\mathfrak{d}^2(\mathfrak{d}-r)^2}\right) \\ &= \tfrac{1}{2\mathfrak{d}^2(\mathfrak{d}-r)}\left(\tfrac{\mathfrak{d}^2r^{-1}-(1+\mathfrak{d}r^{-1})(\mathfrak{d}-r)}{r}\right) = \tfrac{1}{2\mathfrak{d}^2(\mathfrak{d}-r)}. \end{aligned}$$

□[3]

Für die Taylor-Entwicklung sind Ableitungen notwendig, diese ergeben sich für $z \in \Upsilon_{\mathfrak{m}\mathfrak{l}}$ mit der Ketten- und Quotientenregel

$$\begin{aligned} f'_{\|\cdot\|}(z) &= \tfrac{1}{2}\left(-2\tfrac{\langle \mathfrak{l}, \mathfrak{m}-z\mathfrak{l}\rangle_E}{f_{\|\cdot\|}(z)}\right) = -\tfrac{\langle \mathfrak{l}, \mathfrak{m}-z\mathfrak{l}\rangle_E}{f_{\|\cdot\|}(z)} \\ f''_{\|\cdot\|}(z) &= -\left(\tfrac{-\langle \mathfrak{l},\mathfrak{l}\rangle_2 f_{\|\cdot\|}(z) - \langle \mathfrak{l}, \mathfrak{m}-z\mathfrak{l}\rangle_E f'_{\|\cdot\|}(z)}{\big(f_{\|\cdot\|}(z)\big)^2}\right) = \tfrac{\|\mathfrak{l}\|_2^2 (f_{\|\cdot\|}(z))^2 - \langle \mathfrak{l}, \mathfrak{m}-z\mathfrak{l}\rangle_E^2}{(f_{\|\cdot\|}(z))^3}. \end{aligned}$$

Lemma 4.16 (Taylor-Entwicklung)
Seien $\mathfrak{l}, \mathfrak{m}, c \in \mathbb{R}^3$, mit $\mathfrak{l} \neq 0$ und $\|c\|_2 = 1$ gegeben. Weiter seien für ein $\tau \in [-1,1]$, $r \in [0, \mathfrak{d}_\tau)$ sowie $\Upsilon_r(\tau)$ gegeben. Dann gilt für alle $z \in \Upsilon_r(\tau)$

$$f_e(z) = \left\langle \mathfrak{m} - z\mathfrak{l}, \tfrac{\mathfrak{m}-\tau\mathfrak{l}}{\|\mathfrak{m}-\tau\mathfrak{l}\|_2} - c\right\rangle_E + (z-\tau)^2 \int_0^1 \tfrac{\|\mathfrak{l}\|_2^2\|\mathfrak{m}\|_2^2 - \langle \mathfrak{l},\mathfrak{m}\rangle_2^2}{(f_{\|\cdot\|}(\tau+(z-\tau)s))^3}(1-s)\,\mathrm{d}s.$$

Beweis: Nutze für $z \in \Upsilon_r(\tau)$ die folgende Parametrisierung für die Taylor-Entwicklung

$$p_z : [0,1] \to \Upsilon_r(\tau) \subset \Upsilon_{\mathfrak{m}\mathfrak{l}}, \qquad\qquad s \mapsto \tau + (z-\tau)s,$$

für die $p_z'(s) = z - \tau$ für alle $s \in [0,1]$ gilt. Weiter gilt offensichtlich

$$\left(f_{\|\cdot\|} \circ p_z\right)(0) = f_{\|\cdot\|}(\tau) \qquad \text{sowie} \qquad \left(f_{\|\cdot\|} \circ p_z\right)(1) = f_{\|\cdot\|}(z).$$

Die Taylor-Entwicklung von $f_{\|\cdot\|} \circ p_z$ im Punkt $s = 0$ liefert für $f_e(z)$

$$\begin{aligned} f_e(z) &= f_{\|\cdot\|}(z) - \langle \mathfrak{m} - z\mathfrak{l}, c\rangle_E \\ &= (f_{\|\cdot\|} \circ p_z)(0) + (f_{\|\cdot\|} \circ p_z)'(0) + \int_0^1 (f_{\|\cdot\|} \circ p_z)''(s)(1-s)\,\mathrm{d}s - \langle \mathfrak{m} - z\mathfrak{l}, c\rangle_E \\ &= f_{\|\cdot\|}(\tau) + (z-\tau) f_{\|\cdot\|}'(\tau) + (z-\tau)^2 \int_0^1 (f_{\|\cdot\|}'' \circ p_z)(s)(1-s)\,\mathrm{d}s - \langle \mathfrak{m} - z\mathfrak{l}, c\rangle_E. \end{aligned}$$

Das Einsetzen des Funktionsterms für $f_{\|\cdot\|}$ und seiner ersten Ableitung ermöglicht das Zusammenfassen der Nicht-Integralterme

$$\begin{aligned} f_{\|\cdot\|}(\tau) &+ (z-\tau) f_{\|\cdot\|}'(\tau) - \langle \mathfrak{m} - z\mathfrak{l}, c\rangle_E \\ &= \|\mathfrak{m} - \tau\mathfrak{l}\|_2 - (z-\tau)\tfrac{\langle \mathfrak{l}, \mathfrak{m}-\tau\mathfrak{l}\rangle_2}{\|\mathfrak{m}-\tau\mathfrak{l}\|_2} - \langle \mathfrak{m} - z\mathfrak{l}, c\rangle_E \\ &= \tfrac{1}{\|\mathfrak{m}-\tau\mathfrak{l}\|_2}\left(\langle \mathfrak{m}-\tau\mathfrak{l}, \mathfrak{m}-\tau\mathfrak{l}\rangle_2 - \langle \mathfrak{l}(z-\tau), \mathfrak{m}-\tau\mathfrak{l}\rangle_E\right) - \langle \mathfrak{m} - z\mathfrak{l}, c\rangle_E \\ &= \left\langle \mathfrak{m} - z\mathfrak{l}, \tfrac{\mathfrak{m}-\tau\mathfrak{l}}{\|\mathfrak{m}-\tau\mathfrak{l}\|_2}\right\rangle_E - \langle \mathfrak{m} - z\mathfrak{l}, c\rangle_E \\ &= \left\langle \mathfrak{m} - z\mathfrak{l}, \tfrac{\mathfrak{m}-\tau\mathfrak{l}}{\|\mathfrak{m}-\tau\mathfrak{l}\|_2} - c\right\rangle_E. \end{aligned}$$

Im Fall des Integranden ergibt sich durch Einsetzen der zweiten Ableitung

$$\begin{aligned} (f_{\|\cdot\|}'' &\circ p_z)(s) \\ &= \tfrac{\|\mathfrak{l}\|_2^2 \langle \mathfrak{m}-p_z(s)\mathfrak{l}, \mathfrak{m}-p_z(s)\mathfrak{l}\rangle_E - \langle \mathfrak{l}, \mathfrak{m}-p_z(s)\mathfrak{l}\rangle_E^2}{(f_{\|\cdot\|}\circ p_z)^3(s)} \\ &= \tfrac{\|\mathfrak{l}\|_2^2\left(\|\mathfrak{m}\|_2^2 - 2p_z(s)\langle \mathfrak{l}, \mathfrak{m}\rangle_2 + p_z^2(s)\|\mathfrak{l}\|_2^2\right) - \left(\langle \mathfrak{l}, \mathfrak{m}\rangle_2^2 - 2p_z(s)\langle \mathfrak{l}, \mathfrak{m}\rangle_2\|\mathfrak{l}\|_2^2 + p_z^2(s)\|\mathfrak{l}\|_2^4\right)}{(f_{\|\cdot\|}\circ p_z)^3(s)} \\ &= \tfrac{\|\mathfrak{l}\|_2^2\|\mathfrak{m}\|_2^2 - \langle \mathfrak{l}, \mathfrak{m}\rangle_2^2}{(f_{\|\cdot\|}\circ p_z)^3(s)}. \end{aligned}$$

Halte noch die Beobachtung fest, dass damit unabhängig vom Wert von $p_z(s)$ immer

$$\|\mathfrak{l}\|_2^2 \langle \mathfrak{m} - p_z(s)\mathfrak{l}, \mathfrak{m} - p_z(s)\mathfrak{l}\rangle_E - \langle \mathfrak{l}, \mathfrak{m} - p_z(s)\mathfrak{l}\rangle_E^2 = \|\mathfrak{l}\|_2^2\|\mathfrak{m}\|_2^2 - \langle \mathfrak{l}, \mathfrak{m}\rangle_2^2 \tag{4.2.9}$$

gilt. □[3]

Da Abschätzungen für den Betrag des Exponentialterms gesucht werden, verwende, dass für alle $z \in \mathbb{C}$ und $a, b \in \mathbb{R}$ mit $z = a + \mathrm{i}b$

$$|e^z| = |e^{a+\mathrm{i}b}| = |e^a||e^{\mathrm{i}b}| = |e^a| \leq e^{|a|}$$

folgt. Für den Exponentialterm mit der Funktion f_e im Exponenten führt dies zu

$$|e^{\mathrm{i}\kappa(f_{\|\cdot\|}(z) - \langle \mathfrak{m} - z\mathfrak{l}, c\rangle_E)}| = |e^{\mathrm{i}\kappa f_e(z)}| = |e^{-\Im(\kappa f_e(z))}| \leq e^{|\Im(\kappa f_e(z))|}. \tag{4.2.10}$$

Lemma 4.17 (Schranke Exponentenfunktion)
Seien die Voraussetzungen von Lemma 4.16 erfüllt, dann kann der Imaginärteil der Exponentenfunktion für alle $z \in \Upsilon_r(\tau)$ durch

$$|\Im(f_e(z))| \leq \|\mathfrak{l}\|_2 \left(r \left\| \frac{\mathfrak{m}-\tau\mathfrak{l}}{\|\mathfrak{m}-\tau\mathfrak{l}\|_2} - c \right\|_2 + \frac{1}{2(\mathfrak{d}_\tau - r)} r^2 \right)$$

abgeschätzt werden.

Beweis: Vorweg stelle fest, dass für alle $z \in \mathbb{C}$ mit $z = a + \mathrm{i}b$ und $a, b \in \mathbb{R}$

$$z - \tau = (a - \tau) + \mathrm{i}b, \qquad |z - \tau| = \sqrt{(a-\tau)^2 + b^2} \geq \sqrt{b^2} = |b|$$

gilt. Mit der Dreiecksungleichung können die Terme der Taylor-Entwicklung von $f_e(z)$ einzeln abgeschätzt werden. Beginne mit dem Imaginärteil des Nicht-Integralterms aus Lemma 4.16. Für $z \in \Upsilon_r(\tau)$ mit $z = a + \mathrm{i}b$ und der Cauchy-Schwarz-Ungleichung folgt

$$\begin{aligned} \Im\left(\left\langle \mathfrak{m} - z\mathfrak{l}, \frac{\mathfrak{m}-\tau\mathfrak{l}}{\|\mathfrak{m}-\tau\mathfrak{l}\|_2} - c \right\rangle_E \right) &\leq \left| \left\langle \mathfrak{l}b, \frac{\mathfrak{m}-\tau\mathfrak{l}}{\|\mathfrak{m}-\tau\mathfrak{l}\|_2} - c \right\rangle_2 \right| \overset{C.S.}{\leq} \|\mathfrak{l}\|_2 |b| \left\| \frac{\mathfrak{m}-\tau\mathfrak{l}}{\|\mathfrak{m}-\tau\mathfrak{l}\|_2} - c \right\|_2 \\ &\leq \|\mathfrak{l}\|_2 |z-\tau| \left\| \frac{\mathfrak{m}-\tau\mathfrak{l}}{\|\mathfrak{m}-\tau\mathfrak{l}\|_2} - c \right\|_2 \leq \|\mathfrak{l}\|_2 r \left\| \frac{\mathfrak{m}-\tau\mathfrak{l}}{\|\mathfrak{m}-\tau\mathfrak{l}\|_2} - c \right\|_2 . \end{aligned}$$

Der verbleibende Integralterm wird durch Abschätzen seines Betrags beschränkt. Es gilt für alle $s \in [0,1]$ mit $z \in \Upsilon_r(\tau)$

$$|p_z(s) - \tau| = |(z-\tau)s| = |z-\tau| s \leq rs,$$

so dass der Nenner des Integranden wie in Lemma 4.13 nach unten durch

$$\left| (f_{\|\cdot\|} \circ p_z)(s) \right| \geq \|\mathfrak{l}\|_2 (\mathfrak{d}_\tau - rs)$$

abgeschätzt werden kann. Nutze die Beobachtung (4.2.9) um

$$\left| \|\mathfrak{l}\|_2^2 \|\mathfrak{m}\|_2^2 - \langle \mathfrak{l}, \mathfrak{m} \rangle_2^2 \right| = \left| \|\mathfrak{l}\|_2^2 \langle \mathfrak{m} - \tau\mathfrak{l}, \mathfrak{m} - \tau\mathfrak{l} \rangle_2 - \langle \mathfrak{l}, \mathfrak{m} - \tau\mathfrak{l} \rangle_2^2 \right| \leq \|\mathfrak{l}\|_2^2 \|\mathfrak{m} - \tau\mathfrak{l}\|_2^2$$

zu erhalten und setze dies im Integralterm ein. Damit ergibt sich für den Integralterm

$$\begin{aligned} \left| (z-\tau)^2 \int_0^1 \frac{\|\mathfrak{l}\|_2^2 \|\mathfrak{m}\|_2^2 - \langle \mathfrak{l}, \mathfrak{m} \rangle_2^2}{((f_{\|\cdot\|} \circ p_z)(s))^3} (1-s) \, \mathrm{d}s \right| &\overset{\Delta}{\leq} |z-\tau|^2 \int_0^1 \frac{\left| \|\mathfrak{l}\|_2^2 \|\mathfrak{m}\|_2^2 - \langle \mathfrak{l}, \mathfrak{m} \rangle_2^2 \right|}{|((f_{\|\cdot\|} \circ p_z)(s))^3|} (1-s) \, \mathrm{d}s \\ &\leq |z-\tau|^2 \frac{\|\mathfrak{m}-\tau\mathfrak{l}\|_2^2 \|\mathfrak{l}\|_2^2}{\|\mathfrak{l}\|_2^3} \int_0^1 \frac{1-s}{(\mathfrak{d}_\tau - rs)^3} \, \mathrm{d}s \\ &\overset{4.15}{=} |z-\tau|^2 \|\mathfrak{l}\|_2 \frac{\|\mathfrak{m}-\tau\mathfrak{l}\|_2^2}{\|\mathfrak{l}\|_2^2} \frac{1}{2\mathfrak{d}_\tau^2 (\mathfrak{d}_\tau - r)} \\ &= \|\mathfrak{l}\|_2 \frac{\mathfrak{d}_\tau^2}{2\mathfrak{d}_\tau^2 (\mathfrak{d}_\tau - r)} |z-\tau|^2 \\ &= \frac{\|\mathfrak{l}\|_2}{2(\mathfrak{d}_\tau - r)} |z-\tau|^2 \leq \frac{\|\mathfrak{l}\|_2}{2(\mathfrak{d}_\tau - r)} r^2 \end{aligned}$$

wodurch die Behauptung folgt. □[3]

Damit sind alle Abschätzungen beisammen, um den Interpolationsfehler für den Einfachschichtoperator zu beschränken.

4.3 Einfachschichtoperator

Als Zugang zu einer Fehlerschranke wird der Satz zur holomorphen Approximation 4.8 verwendet. Hierzu gilt es, eine Abschätzung für das dort auftretende Maximum der approximierten Funktion auf der Bernstein-Ellipse zu finden. Im Anschluss wird dieses Resultat mit Hilfe einer Bestapproximationsaussage auf den zu untersuchenden Interpolationsfehler übertragen. Der Beweis wurde abgewandelt und stammt ursprünglich aus [3].

Lemma 4.18 (Approximation)
Seien ein richtungsabhängiger Blockbaum $\mathcal{T}_{\mathcal{I}\times\mathcal{I}}$ sowie ein zulässiges Blatt $b = (t, s, c) \in \mathcal{T}_{\mathcal{I}\times\mathcal{I}}$ und ein $i \in \underline{6}$ gegeben. Weiter seien $\mathfrak{m}, \mathfrak{l}$ und $\mathfrak{d}$ passend zur Interpolation gegeben. Setze für $r \in (0, \mathfrak{d})$ den Halbachsenparameter der Bernstein-Ellipse mit $\varrho = \sqrt{r^2+1} + r$. Dann existiert zu der Interpolationsordnung $m \in \mathbb{N}$ ein Polynom $p \in \Pi_m$, so dass

$$\left\| {}^i g_{ec} - p \right\|_{\infty,[-1,1]} \leq \tfrac{2}{\operatorname{dist}(Q_t, Q_s)} \varrho^{-m} \tfrac{\mathcal{C}_e^\star(\eta_1,\eta_2,r)}{r}$$

gilt, wobei $\mathcal{C}_e^\star(\eta_1, \eta_2, r)$ durch

$$\mathcal{C}_e^\star(\eta_1, \eta_2, r) := \frac{e^{\frac{1}{2}\left(r(\eta_1+\eta_2)+\frac{\eta_2}{4(1-\frac{r}{\mathfrak{d}})}r^2\right)}}{4\pi(1-\frac{r}{\mathfrak{d}})}$$

definiert ist.

Beweis: Da ${}^i g_{ec}$ auf der Bernstein-Ellipse D_ϱ holomorph ist, existiert nach Satz 4.8 ein Polynom $p \in \Pi_m$, so dass mit Bemerkung 20

$$\left\| {}^i g_{ec} - p \right\|_{\infty,[-1,1]} \leq \tfrac{2}{(\varrho-1)} \varrho^{-m} \max\left\{ |{}^i g_{ec}(z)| \mid z \in \overline{D}_\varrho \right\}$$

gilt. Als Erstes schätze den Nenner $\varrho - 1$ ab

$$\varrho - 1 = \sqrt{r^2+1} + r - 1 \geq r,$$

um damit für den Bruch

$$\tfrac{2}{(\varrho-1)} \leq \tfrac{2}{r}$$

zu erhalten.
Um das Maximum von ${}^i g_{ec}$ auf dem Rand der Bernstein-Ellipse D_ϱ abzuschätzen, nutze die Eigenschaften von $\mathfrak{m}, \mathfrak{l}$ nach (4.2.2)

$$\mathfrak{m} - \tau\mathfrak{l} \in \{\widetilde{x} - \widetilde{y} \mid \widetilde{x} \in Q_t, \widetilde{y} \in Q_s\} = Q_t - Q_s \qquad \text{für alle } \tau \in [-1, 1]$$

sowie

$$\|\mathfrak{l}\|_2 \leq \tfrac{\max\{\operatorname{diam}(Q_t), \operatorname{diam}(Q_s)\}}{2}.$$

Dies zusammen mit der Aussage zur Richtung (2.3.1)

$$\kappa \left\| \frac{x-y}{\|x-y\|_2} - c \right\|_2 \leq \frac{\eta_1+\eta_2}{\max\{\operatorname{diam}(Q_t),\operatorname{diam}(Q_s)\}}$$

führt zu

$$\left\| \frac{\mathfrak{m}-\tau\mathfrak{l}}{\|\mathfrak{m}-\tau\mathfrak{l}\|_2} - c \right\|_2 \leq \frac{\eta_1+\eta_2}{\kappa \max\{\operatorname{diam}(Q_t),\operatorname{diam}(Q_s)\}} \leq \frac{\eta_1+\eta_2}{2\kappa\|\mathfrak{l}\|_2} \qquad \text{für alle } \tau \in [-1,1].$$

Mit Bemerkung 21 zusammen kann dies genutzt werden, um die Aussage zum Exponenten aus Lemma 4.17 weiter zu verfeinern

$$\begin{aligned}
|\Im(f_e(z))| &\leq \|\mathfrak{l}\|_2 \left(r \left\| \frac{\mathfrak{m}-\tau\mathfrak{l}}{\|\mathfrak{m}-\tau\mathfrak{l}\|_2} - c \right\|_2 + \frac{1}{2(\mathfrak{d}_\tau - r)} r^2 \right) \\
&\leq \|\mathfrak{l}\|_2 \left(r \frac{\eta_1+\eta_2}{2\kappa\|\mathfrak{l}\|_2} + \frac{1}{2\mathfrak{d}(1-\frac{r}{\mathfrak{d}})} r^2 \right) \\
&\leq \|\mathfrak{l}\|_2 \left(r \frac{\eta_1+\eta_2}{2\kappa\|\mathfrak{l}\|_2} + \frac{\eta_2}{8\kappa\|\mathfrak{l}\|_2(1-\frac{r}{\mathfrak{d}})} r^2 \right) \\
&= \frac{1}{2\kappa} \left(r(\eta_1+\eta_2) + \frac{\eta_2}{4(1-\frac{r}{\mathfrak{d}})} r^2 \right),
\end{aligned}$$

was wiederum der Abschätzung des Zählers dient.
Um eine Schranke für den Nenner zu finden, nutze erneut Bemerkung 21 zusammen mit Lemma 4.13

$$\begin{aligned}
|f_{\|\cdot\|}(z)| &\geq \|\mathfrak{l}\|_2(\mathfrak{d}-r) = \|\mathfrak{l}\|_2\mathfrak{d}(1-\tfrac{r}{\mathfrak{d}}) \geq \|\mathfrak{l}\|_2 \frac{\operatorname{dist}(Q_t,Q_s)}{\|\mathfrak{l}\|_2}(1-\tfrac{r}{\mathfrak{d}}) \\
&= \operatorname{dist}(Q_t,Q_s)(1-\tfrac{r}{\mathfrak{d}}).
\end{aligned}$$

Das Zusammenfügen dieser Ergebnisse liefert eine Abschätzung für die Funktion ${}^i g_{ec}$ und damit auch für ihr Maximum für $z \in \overline{D}_\varrho$

$$\left|{}^i g_{ec}(z)\right| \leq \frac{e^{|\kappa\Im(f_e(z))|}}{4\pi|f_{\|\cdot\|}(z)|} \leq \frac{e^{\frac{1}{2}\left(r(\eta_1+\eta_2)+\frac{\eta_2}{4(1-\frac{r}{\mathfrak{d}})}r^2\right)}}{4\pi \operatorname{dist}(Q_t,Q_s)(1-\frac{r}{\mathfrak{d}})},$$

was den Beweis mit

$$\left\|{}^i g_{ec} - p\right\|_{\infty,[-1,1]} \leq \frac{2}{\operatorname{dist}(Q_t,Q_s)} \varrho^{-m} \frac{1}{r} \underbrace{\frac{e^{\frac{1}{2}\left(r(\eta_1+\eta_2)+\frac{\eta_2}{4(1-\frac{r}{\mathfrak{d}})}r^2\right)}}{4\pi(1-\frac{r}{\mathfrak{d}})}}_{=\mathcal{C}_e^\star(\eta_1,\eta_2,r)},$$

vervollständigt. □

Insgesamt führt dies zu einer Abschätzung für den Interpolationsfehler, die erstmals in leicht abgewandelter Form in [3] zu finden ist.

Theorem 4.19 (Approximationsfehler)
Seien ein richtungsabhängiger Blockbaum $\mathcal{T}_{\mathcal{I}\times\mathcal{I}}$ sowie ein zulässiges Blatt $b = (t, s, c) \in \mathcal{T}_{\mathcal{I}\times\mathcal{I}}$ gegeben. Weiter seien passend zur Interpolation $m \in \mathbb{N}_0$ sowie $\mathfrak{d}$, ein $r \in (0, \mathfrak{d})$ und $\varrho = \sqrt{r^2+1} + r$ gegeben, dann ist der Interpolationsfehler beschränkt durch

$$\left\| g_e - \mathfrak{I}^c_{Q_t\times Q_s}[g_e] \right\|_{\infty, Q_t\times Q_s} \leq (1+\Lambda_m)\tfrac{12\Lambda_m^5}{\operatorname{dist}(Q_t,Q_s)}\varrho^{-m}\tfrac{\mathcal{C}_e^\star(\eta_1,\eta_2,r)}{r},$$

mit $\mathcal{C}_e^\star(\eta_1, \eta_2, r)$ wie in Lemma 4.18.

Beweis: Auf Grund von (4.2.1) ist es möglich, die Analyse auf

$$\|g_{ec} - \mathfrak{I}_{Q_t\times Q_s}[g_{ec}]\|_{\infty, Q_t\times Q_s}$$

zu beschränken. Um das Lemma 2.7 anwenden zu können, ist es nötig, eine Abschätzung für

$$\left\| {}^i g_{ec} - \mathfrak{I}_{[-1,1]}[{}^i g_{ec}] \right\|_{\infty,[-1,1]} \qquad \text{für alle } i \in \underline{6}$$

zu finden.
Da der Interpolationsoperator m-ten Grades Polynome vom Grad kleiner oder gleich m unverändert lässt, folgt für das Polynom $p \in \Pi_m$ aus Satz 4.8

$$\begin{aligned}\left\| {}^i g_{ec} - \mathfrak{I}_{[-1,1]}[{}^i g_{ec}] \right\|_{\infty,[-1,1]} &= \left\| {}^i g_{ec} - p - \mathfrak{I}_{[-1,1]}[{}^i g_{ec} - p] \right\|_{\infty,[-1,1]} \\ &\overset{\Delta}{\leq} \left\| {}^i g_{ec} - p \right\|_{\infty,[-1,1]} + \left\| \mathfrak{I}_{[-1,1]}[{}^i g_{ec} - p] \right\|_{\infty,[-1,1]} \\ &\leq (1+\Lambda_m) \left\| {}^i g_{ec} - p \right\|_{\infty,[-1,1]}.\end{aligned}$$

Die verbliebene Norm kann mit Hilfe des Lemmas 4.18 abgeschätzt werden. Es folgt

$$\left\| {}^i g_{ec} - \mathfrak{I}_{[-1,1]}[{}^i g_{ec}] \right\|_{\infty,[-1,1]} \leq (1+\Lambda_m)\tfrac{2}{\operatorname{dist}(Q_t,Q_s)}\varrho^{-m}\tfrac{\mathcal{C}_e^\star(\eta_1,\eta_2,r)}{r}$$

und mit Lemma 2.7 die Behauptung. □

In Anlehnung an die Ergebnisse in [3] kann die Aussage zum Approximationsfehler unter zusätzlichen Annahmen noch vereinfacht werden.

Korollar 4.20
Seien die Voraussetzungen von Theorem 4.19 erfüllt und es gelte $r = \min\left\{1, \frac{3}{4}\mathfrak{d}\right\}$, dann kann eine knappere und damit handlichere Fehlerdarstellung

$$\left\| g_e - \mathfrak{I}^c_{Q_t\times Q_s}[g_e] \right\|_{\infty, Q_t\times Q_s} \leq (1+\Lambda_m) 12\Lambda_m^5 \varrho^{-m} \tfrac{\mathcal{C}_{\eta_2} e^{\eta_1+\eta_2}}{\pi \operatorname{dist}(Q_t,Q_s)}$$

mit $\mathcal{C}_{\eta_2} := \max\left\{1, \frac{2\eta_2}{3}\right\}$ erhalten werden.

Beweis: Mit Bemerkung 21 gilt

$$\tfrac{3}{4}\mathfrak{d} \geq \tfrac{3}{4}\tfrac{2}{\eta_2} = \tfrac{3}{2\eta_2},$$

was zu $r \geq \min\left\{1, \frac{3}{2\eta_2}\right\}$ führt. Damit kann der Nenner von $\mathcal{C}_e^\star(\eta_1, \eta_2, r)r^{-1}$ noch weiter vereinfacht werden

$$4\pi r(1 - \tfrac{r}{\mathfrak{d}}) \geq 4\pi \min\left\{1, \tfrac{3}{2\eta_2}\right\}(1 - \tfrac{3\mathfrak{d}}{4\mathfrak{d}}) \geq 4\pi\tfrac{1}{4}\min\left\{1, \tfrac{3}{2\eta_2}\right\} = \pi\min\left\{1, \tfrac{3}{2\eta_2}\right\}.$$

Außerdem lässt sich das Argument der Exponentialfunktion im Zähler von $\mathcal{C}_e^\star(\eta_1, \eta_2, r)$ weiter abschätzen mit

$$\begin{aligned}\tfrac{1}{2}\left(r(\eta_1+\eta_2) + \tfrac{\eta_2}{4(1-\frac{r}{\mathfrak{d}})}r^2\right) &\leq \tfrac{1}{2}\left(\eta_1+\eta_2+\tfrac{\eta_2}{4(1-\frac{r}{\mathfrak{d}})}\right) \leq \tfrac{1}{2}(\eta_1+2\eta_2)\\ &\leq \eta_1+\eta_2,\end{aligned}$$

was zur Behauptung führt. □

4.3.1 Reinterpolation

Eine Abschätzung für den Fehler der direkten Interpolation allein reicht nicht aus. Durch die Eigenschaft der Schachtelung der Clusterbasen kommt beim Wechsel der Stufe eine Reinterpolation hinzu (siehe Kapitel 3.1.1). Teilen sich die beiden Stufen nicht dieselben Richtungen, so führt dies zu einem zusätzlichen Fehler. Ziel dieses Abschnitts ist es, ähnlich wie in [3, Kap. 5] zu zeigen, dass auch dieser Fehler kontrollierbar ist.

Zum Erreichen dieses Ziels gilt es zunächst, Notationen für die auftretenden Sequenzen von Clustern und den dazugehörigen Richtungen einzuführen.
Für $t, s \in \mathcal{T}_\mathcal{I}$ und ein passendes $L \in \underline{p_{\mathcal{I}\rfloor}}$ bezeichne mit $\mathfrak{s}_t^L$ und $\mathfrak{s}_s^L$ durch Pfade im Clusterbaum definierte Tupel von Clustern

$$\begin{aligned}\mathfrak{s}_t^L &= (t_0, t_1, \ldots, t_L) && \text{mit } t = t_0, t_i \in \operatorname{kind}(t_{i-1}) \text{ für alle } i \in \underline{L\rfloor},\\ \mathfrak{s}_s^L &= (s_0, s_1, \ldots, s_L) && \text{mit } s = s_0, s_i \in \operatorname{kind}(s_{i-1}) \text{ für alle } i \in \underline{L\rfloor},\end{aligned}$$

deren korrespondierende überdeckende Quader

$$Q_{t_0} \supseteq Q_{t_1} \supseteq \cdots \supseteq Q_{t_L} \qquad \text{beziehungsweise} \qquad Q_{s_0} \supseteq Q_{s_1} \supseteq \cdots \supseteq Q_{s_L}$$

erfüllen. Eine zu $\mathfrak{s}_t^L$ passende Sequenz an Richtungen wird mit $\mathfrak{s}_c^L$ bezeichnet

$$\mathfrak{s}_c^L = (c_0, c_1, \ldots, c_L) \qquad \text{mit } c_i \in \mathcal{R}_{t_i} \text{ für alle } i \in \underline{L\rfloor}_0.$$

Mit diesen Sequenzen an Clustern und Richtungen kann der Pfad im Clusterbaum, entlang dem eine Rekonstruktion der Clusterbasis notwendig wird, verfolgt und damit die nötige

Reinterpolation untersucht werden.
Auch die Notation des Interpolationsoperators soll an die in jeder Stufe möglicherweise wechselnde ebene Welle angepasst werden. Die Interpolation auf dem Block $(t_\ell, s_\ell, c_\ell) \in \mathcal{T}_{\mathcal{I}\times\mathcal{I}}$ ist ebenfalls als Tensorprodukt gegeben

$$\mathfrak{I}^{c_\ell}_{Q_{t_\ell}\times Q_{s_\ell}} = \mathfrak{I}^{c_\ell}_{Q_{t_\ell}} \otimes \mathfrak{I}^{-c_\ell}_{Q_{s_\ell}},$$

wobei die Definition des Operators $\mathfrak{I}^{c_\ell}_{Q_{t_\ell}}$ sich an der Definition 3.1.1 orientiert. Er ist durch

$$\mathfrak{I}^{c_\ell}_{Q_{t_\ell}}[u](x) = e^{\mathrm{i}\kappa\langle x, c_\ell\rangle_2}\mathfrak{I}_{Q_{t_\ell}}[e^{-\mathrm{i}\kappa\langle x, c_\ell\rangle_2}u](x) \qquad \text{für alle } \ell \in L\rfloor_0, u \in C(Q_{t_\ell}), x \in Q_{t_\ell}$$

gegeben, analog verfahre mit dem Interpolationsoperator $\mathfrak{I}^{-c_\ell}_{Q_{s_\ell}}$ für die Spaltencluster. Im Kapitel 4.3 wurde der Fehler auf dem Block $(t, s, c) \in \mathcal{L}^+_{\mathcal{I}\times\mathcal{I}}$ für alle $x \in Q_t, y \in Q_s$, also die Differenz

$$g_{ec}(x,y) - \mathfrak{I}_{Q_t\times Q_s}[g_{ec}](x,y) = g_{ec}(x,y) - \sum_{\widehat{\mu},\widehat{\nu}\in\widehat{M}} g_{ec}(x_{\widehat{\mu}}, y_{\widehat{\nu}})\ell_{t,\widehat{\mu}}(x)\overline{\ell_{s,\widehat{\nu}}(y)},$$

untersucht. Ist für einen gegebenen Block (t, s, c) mindestens einer der Cluster t oder s noch kein Blatt, ist eine weitere Untersuchung notwendig. Seien $L \in \underline{p_{\mathcal{I}}}\rfloor$, die maximale Anzahl an Stufen bis die Nachfahren von t und s Blätter sind, $\mathfrak{s}^L_t$, $\mathfrak{s}^L_s$ die dazugehörigen Clustersequenzen und Richtungen $\mathfrak{s}^L_c$, die $(t_0, s_0, c_0) = (t, s, c)$ erfüllen. Es gilt dann noch,

$$\begin{aligned} g_e(x,y) - &\mathfrak{I}^{c_L}_{Q_{t_L}\times Q_{s_L}} \circ \cdots \circ \mathfrak{I}^{c_0}_{Q_{t_0}\times Q_{s_0}}[g_e](x,y) \\ &= g_e(x,y) - \sum_{\widehat{\mu},\widehat{\nu}\in\widehat{M}} g_{ec}(x_{\widehat{\mu}}, y_{\widehat{\nu}})\ell^{\mathfrak{s}^L_t,\mathfrak{s}^L_c}_{\widehat{\mu}}(x)\overline{\ell^{\mathfrak{s}^L_s,\mathfrak{s}^L_c}_{\widehat{\nu}}(y)} \end{aligned} \tag{4.3.1}$$

zu analysieren, wobei die reinterpolierten Lagrange-Polynome mit Interpolationspunkten $\{x_{\widehat{\mu}}\}_{\widehat{\mu}\in\widehat{M}} \subset Q_{t_0}$ durch

$$\ell^{\mathfrak{s}^L_t,\mathfrak{s}^L_c}_{\widehat{\mu}}(x) := \mathfrak{I}^{c_L}_{Q_{t_L}} \circ \cdots \circ \mathfrak{I}^{c_1}_{Q_{t_1}}[\ell_{t_0 c_0,\widehat{\mu}}](x) \qquad \text{für alle } \widehat{\mu} \in \widehat{M}, x \in Q_{t_L} \tag{4.3.2}$$

gegeben sind. Falls die Pfade zu den Blattnachfahren von s und t unterschiedlich lang sind, definiere die zusätzlichen Interpolationen beim kürzeren Pfad als Identitäten.
Die Differenz (4.3.1) kann mit Hilfe einer Teleskopsumme zu

$$\sum_{\ell=0}^{L} \mathfrak{I}^{c_L}_{Q_{t_L}\times Q_{s_L}} \circ \cdots \circ \mathfrak{I}^{c_{\ell+1}}_{Q_{t_{\ell+1}}\times Q_{s_{\ell+1}}}[g_e - \mathfrak{I}^{c_\ell}_{Q_{t_\ell}\times Q_{s_\ell}}[g_e]],$$

umformuliert werden und lässt sich entsprechend durch

$$\begin{aligned} &\sum_{\ell=0}^{L} \left\| \mathfrak{I}^{c_L}_{Q_{t_L}\times Q_{s_L}} \circ \cdots \circ \mathfrak{I}^{c_{\ell+1}}_{Q_{t_{\ell+1}}\times Q_{s_{\ell+1}}} \right\|_{op, C(Q_{t_L}\times Q_{s_L})\leftarrow C(Q_{t_\ell}\times Q_{s_\ell})} \\ &\cdot \left\| g_e - \mathfrak{I}^{c_\ell}_{Q_{t_\ell}\times Q_{s_\ell}}[g_e] \right\|_{\infty, Q_{t_\ell}\times Q_{s_\ell}} \end{aligned}$$

beschränken.
Auch der geschachtelte richtungsabhängige Interpolationsoperator kann über die Eigenschaften des Tensorprodukts wieder mit Operatoren auf eindimensionalen Gebieten geschrieben werden. Zunächst gilt

$$\mathfrak{I}^{c_L}_{Q_{t_L}\times Q_{s_L}} \circ \cdots \circ \mathfrak{I}^{c_{\ell+1}}_{Q_{t_{\ell+1}}\times Q_{s_{\ell+1}}} = \left(\mathfrak{I}^{c_L}_{Q_{t_L}} \circ \cdots \circ \mathfrak{I}^{c_{\ell+1}}_{Q_{t_{\ell+1}}}\right) \otimes \left(\mathfrak{I}^{-c_L}_{Q_{s_L}} \circ \cdots \circ \mathfrak{I}^{-c_{\ell+1}}_{Q_{s_{\ell+1}}}\right),$$

so dass für die Analyse der Stabilität

$$\begin{aligned}&\left\|\mathfrak{I}^{c_L}_{Q_{t_L}\times Q_{s_L}} \circ \cdots \circ \mathfrak{I}^{c_{\ell+1}}_{Q_{t_{\ell+1}}\times Q_{s_{\ell+1}}}\right\|_{op,C(Q_{t_L}\times Q_{s_L})\leftarrow C(Q_{t_\ell}\times Q_{s_\ell})}\\ &\leq \left\|\mathfrak{I}^{c_L}_{Q_{t_L}} \circ \cdots \circ \mathfrak{I}^{c_{\ell+1}}_{Q_{t_{\ell+1}}}\right\|_{op,C(Q_{t_L})\leftarrow C(Q_{t_\ell})} \left\|\mathfrak{I}^{-c_L}_{Q_{s_L}} \circ \cdots \circ \mathfrak{I}^{-c_{\ell+1}}_{Q_{s_{\ell+1}}}\right\|_{op,C(Q_{s_L})\leftarrow C(Q_{s_\ell})}\end{aligned} \tag{4.3.3}$$

zu untersuchen ist. Da die Cluster t und s voneinander getrennt untersucht werden können, ist es auch möglich, mit unterschiedlichen Pfadlängen L zu arbeiten. Bei $\mathfrak{I}^{c_\ell}_{Q_{t_\ell}}$ für $\ell \in \underline{L}_\rfloor$ handelt es sich ebenfalls um Tensorprodukte von Operatoren. Da der Exponentialterm

$$e^{\mathrm{i}\kappa\langle x,c_\ell\rangle_2} = e^{\mathrm{i}\kappa x_1(c_\ell)_1} e^{\mathrm{i}\kappa x_2(c_\ell)_2} e^{\mathrm{i}\kappa x_3(c_\ell)_3} \qquad \text{für alle } x \in Q_{t_\ell}$$

erfüllt, ist die Analyse auf den eindimensionalen Fall zurückführbar, was auch die Hauptarbeit auf den eindimensionalen Fall beschränkt.
Seien Sequenzen an Clustern $\mathfrak{s}^L_t$, $\mathfrak{s}^L_s$ gegeben und bezeichne mit $I_i = \{I_{\ell,i}\}_{\ell\in\underline{L}_\rfloor 0}$ die in der Koordinate $i \in \underline{6}_\rfloor$ zu betrachtende Sequenz von Intervallen. Das Intervall $I_{\ell,i}$ ist durch

$$I_{\ell,i} := \begin{cases}[a_{t_\ell,i}, b_{t_\ell,i}] & \text{falls } i \in \underline{3}_\rfloor,\\ [a_{s_\ell,i-3}, b_{s_\ell,i-3}] & \text{falls } i \in \underline{6}_\rfloor 4\end{cases} \qquad \text{für alle } \ell \in \underline{L}_\rfloor 0 \tag{4.3.4}$$

gegeben, wobei aus der Schachtelung der überdeckenden Quader unmittelbar

$$I_{0,i} \supseteq I_{1,i} \supseteq \cdots \supseteq I_{L,i}$$

folgt. Falls aus dem Kontext heraus klar ist, welche Koordinate $i \in \underline{6}_\rfloor$ betrachtet wird, nutze die Notationen I_ℓ statt $I_{\ell,i}$ und I statt I_i. Notiere die Komponente des zu $\ell \in \underline{L}_\rfloor 0$ und i dazugehörigen Richtungsvektors c_ℓ aus der Richtungssequenz $\mathfrak{s}^L_c$ mit

$$\dot{c}^i_\ell := \begin{cases}(c_\ell)_i & \text{falls } i \in \underline{3}_\rfloor,\\ -(c_\ell)_{i-3} & \text{falls } i \in \underline{6}_\rfloor 4.\end{cases}$$

Schreibe den eindimensionalen angepassten Interpolationsoperator für alle $\ell \in \underline{L}_\rfloor 0$ mit

$$\mathfrak{I}^{i,c_\ell}_{I_\ell}[u](x) = e^{\mathrm{i}\kappa x\dot{c}^i_\ell}\mathfrak{I}_{I_\ell}[e^{-\mathrm{i}\kappa x\dot{c}^i_\ell}u](x) \qquad \text{für alle } i \in \underline{6}_\rfloor, u \in C(I_\ell), x \in I_\ell.$$

Um die Hintereinanderausführung in eine kompaktere Schreibweise zu bringen, führe zu der Intervallsequenz I_i und der Sequenz an Richtungen $\mathfrak{s}_c^L$ den Operator $\mathfrak{I}_I^{i,\mathfrak{s}_c^L}$ mit

$$\mathfrak{I}_I^{i,\mathfrak{s}_c^L} := \mathfrak{I}_{I_L}^{i,c_L} \circ \cdots \circ \mathfrak{I}_{I_1}^{i,c_1} \qquad \text{für alle } i \in \underline{6}$$

ein.
Damit überhaupt eine Beschränktheit des Fehlers vorliegen kann, ist eine Kontraktionseigenschaft der beteiligten Gebiete notwendig. Nehme an, dass ein positives $\mathcal{C}_k \in \mathbb{R}_{<1}$ mit

$$\frac{|I_{\ell,i}|}{|I_{\ell-1,i}|} \leq \mathcal{C}_k \qquad \text{für alle } \ell \in L, i \in \underline{6} \tag{4.3.5}$$

existiert.

Bemerkung 22 (Kontraktionseigenschaft und Clusterstrategien): *Abhängig von der angewendeten Konstruktion der überdeckenden Quader sind nicht alle Clusterstrategien geeignet, um die Kontraktionseigenschaft zu erfüllen. Bei der* adaptiven Clusterstrategie *(Bem. 11) wird pro Schritt nur entlang einer Koordinate j unterteilt, was schnell zu $I_{\ell,i} = I_{\ell-1,i}$ für $i \neq j$ führen kann.*

Im Zuge der angestrebten Abschätzungen ist eine nähere Betrachtung von transformierten Bernstein-Ellipsen notwendig. Auch das Arbeit mit transformierten Bernstein-Ellipsen wird durch die Darstellung mit Hilfe der Joukowski-Transformation erleichtert, wie Herr Börm in einer unveröffentlichten Arbeit zur iterierten Interpolation zeigt [9] und hier Anwendung findet.

Lemma 4.21 (Transformierte Bernstein-Ellipse)
Seien $-1 \leq a < b \leq 1$ und $h := \frac{(b-a)}{2}$ sowie $m := \frac{b+a}{2}$ gegeben. Bezeichne durch

$${}^{a,b}D_\varrho := m + hD_\varrho \qquad \text{für alle } \varrho > 1,$$

die von $[-1,1]$ auf $[a,b]$ transformierte Bernstein-Ellipse D_ϱ. Für alle $\varrho \in \mathbb{R}_{>1}$ und $\widehat{\varrho} \leq \varpi^\dagger\left(\frac{\varpi(\varrho)-(1-h)}{h}\right)$, wobei ϖ die Joukowski-Transformation 4.3 sei, gilt

$${}^{a,b}D_{\widehat{\varrho}} \subset D_\varrho.$$

Beweis: Sei $\varrho > 1$ gegeben. Mit Bemerkung 18 reicht es, für $\widehat{\varrho} = \varpi^\dagger\left(\frac{\varpi(\varrho)-(1-h)}{h}\right)$ die Inklusion der Ellipsen zu zeigen.
Seien zu $\varrho, \widehat{\varrho}$ durch $a_\varrho = \varpi(\varrho)$ und $a_{\widehat{\varrho}} = \varpi(\widehat{\varrho})$ die reellen Halbachsen der zugehörigen Bernstein-Ellipsen gegeben. Für jedes $z \in {}^{a,b}D_{\widehat{\varrho}}$ ist Folgendes äquivalent

$$z \in {}^{a,b}D_{\widehat{\varrho}} \quad \Longleftrightarrow \quad z - m \in hD_{\widehat{\varrho}} \quad \Longleftrightarrow \quad \tfrac{z-m}{h} \in D_{\widehat{\varrho}}.$$

Dies bedeutet, dass folgende Ungleichung erfüllt sein muss

$$\left|\tfrac{z-m}{h}+1\right|+\left|\tfrac{z-m}{h}-1\right| \leq 2a_{\widehat{\varrho}} \iff |z-m+h|+|z-m-h| \leq 2ha_{\widehat{\varrho}},$$

welche mit

$$\begin{aligned} -m+h &= -\tfrac{b+a}{2}+\tfrac{b-a}{2} = -a \\ -m-h &= -\tfrac{b+a}{2}-\tfrac{b-a}{2} = -b \end{aligned}$$

vereinfacht wird zu

$$|z-a|+|z-b| \leq 2ha_{\widehat{\varrho}}.$$

Weiterhin gilt für solch ein z auch

$$\begin{aligned} |z+1|+|z-1| &= |z-a+a+1|+|z-b+b-1| \\ &\overset{\Delta}{\leq} |z-a|+|a+1|+|z-b|+|1-b| \\ &\leq 2ha_{\widehat{\varrho}}+a+2-b \\ &= 2ha_{\widehat{\varrho}}+2-(b-a) \\ &= 2(ha_{\widehat{\varrho}}+1-h). \end{aligned}$$

Damit auch $z \in D_\varrho$ gilt, muss entsprechend

$$2(ha_{\widehat{\varrho}}+1-h) \leq 2a_\varrho$$

erfüllt sein. Dies ist äquivalent zu

$$a_{\widehat{\varrho}} \leq \tfrac{a_\varrho-(1-h)}{h} \iff \varpi(\widehat{\varrho}) \leq \tfrac{\varpi(\varrho)-(1-h)}{h}.$$

Aus $\varrho > 1$ und dem monotonen Wachstum (4.1.4) der Joukowski-Transformation ϖ folgt

$$\tfrac{\varpi(\varrho)-(1-h)}{h} \geq \tfrac{\varpi(1)-(1-h)}{h} = \tfrac{h}{h} = 1,$$

weshalb $\varpi^\dagger$ mit (4.1.5) im Fall der Gleichheit auf beide Seiten angewendet werden kann. Es ergibt sich

$$\widehat{\varrho} = \varpi^\dagger\left(\varpi(\widehat{\varrho})\right) \leq \varpi^\dagger\left(\tfrac{\varpi(\varrho)-(1-h)}{h}\right)$$

und damit die Behauptung. □

Transformierte Bernstein-Ellipsen können auch mit Hilfe einer Umparametrisierung Φ beschrieben werden. Definiere dazu für das Intervall $[a,b]$ die Funktion $\Phi_{[a,b]}$ mit

$$\Phi_{[a,b]} : \mathbb{C} \to \mathbb{C}, \qquad z \mapsto \tfrac{b+a}{2}+\tfrac{b-a}{2}z. \tag{4.3.6}$$

Dann gilt offensichtlich $\Phi_{[a,b]}(D_\varrho) = {}^{a,b}D_\varrho$ für alle $\varrho \in \mathbb{R}_{>1}$. Für $\ell \in \underline{L_\rfloor}_0$, $i \in \underline{6_\rfloor}$ und das dazugehörige Intervall $I_{\ell,i}$ nutze auch die Schreibweise ${}^{I_{\ell,i}}D_\varrho$ für die auf $I_{\ell,i}$ transformierte Bernstein-Ellipse.

Mit der folgenden Hilfsaussage kann die Inklusion der transformierten Bernstein-Ellipsen noch erweitert werden, so dass es möglich wird, Bernstein-Ellipsen zu schachteln. Dazu wird die Schranke für $\widehat{\varrho}$ weiter untersucht. Um dies losgelöst von einem konkreten Intervall und damit von h durchzuführen, nutze aus, dass h unabhängig vom betrachteten Intervall $[a,b] \subset [-1,1]$ immer $h < 1$ erfüllt. Beide folgenden Lemmata stammen ebenfalls aus einer unveröffentlichten Arbeit von Herrn Börm [9].

Lemma 4.22
Sei ein $\gamma \in (0,1)$ gegeben. Definiere die Funktion ϑ durch

$$\vartheta : \mathbb{R}_{\geq 1} \to \mathbb{R}_{\geq 1}, \qquad \alpha \mapsto \frac{\varpi^\dagger\left(\frac{\varpi(\alpha)-1}{\gamma}+1\right)}{\alpha}.$$

Auf $\mathbb{R}_{>1}$ ist ϑ eine streng monoton wachsende Funktion mit den Grenzwerten $\vartheta(1) = 1$ und $\lim_{\alpha\to\infty} \vartheta(\alpha) = \frac{1}{\gamma}$.

Beweis: Um das streng monoton wachsende Verhalten der Funktion zu zeigen, definiere zunächst eine weitere Funktion

$$g : \mathbb{R}_{\geq 1} \to \mathbb{R}_{\geq 1}, \qquad \alpha \mapsto \frac{\varpi^\dagger\left(\gamma^{-1}(\alpha-1)+1\right)}{\varpi^\dagger(\alpha)}.$$

Dann lässt sich ϑ durch $\vartheta = g \circ \varpi$ darstellen. Von der Funktion ϖ ist schon bekannt, dass sie streng monoton wachsend ist, entsprechend reicht es aus, zu zeigen, dass g ebenfalls diese Eigenschaft aufweist. Damit die anschließenden Argumentationen übersichtlich bleiben, nutze Kurzschreibweisen für Zähler und Nenner der Funktion g

$$\begin{aligned} g_z(\alpha) &:= \varpi^\dagger\left(\gamma^{-1}(\alpha-1)+1\right), \\ g_n(\alpha) &:= \varpi^\dagger(\alpha). \end{aligned}$$

Für die Monotonie zeige, dass für alle $\alpha \in \mathbb{R}_{>1}$, $g'(\alpha) > 0$ gilt. Mit der Eigenschaft (4.1.5) der Joukowski-Transformation folgt

$$\tfrac{\mathrm{d}}{\mathrm{d}\alpha} g_n(\alpha) = 1 + \frac{\alpha}{\sqrt{\alpha^2-1}} = \frac{\alpha+\sqrt{\alpha^2-1}}{\sqrt{\alpha^2-1}} = \frac{g_n(\alpha)}{\sqrt{\alpha^2-1}}$$

sowie

$$\begin{aligned} \tfrac{\mathrm{d}}{\mathrm{d}\alpha} g_z(\alpha) &= \frac{1}{\gamma} + \frac{\gamma^{-1}(\alpha-1)+1}{\gamma\sqrt{(\gamma^{-1}(\alpha-1)+1)^2-1}} = \frac{\sqrt{(\gamma^{-1}(\alpha-1)+1)^2-1}+\gamma^{-1}(\alpha-1)+1}{\gamma\sqrt{(\gamma^{-1}(\alpha-1)+1)^2-1}} \\ &= \frac{g_z(\alpha)}{\gamma\sqrt{(\gamma^{-1}(\alpha-1)+1)^2-1}}. \end{aligned}$$

Mit der Quotientenregel ergibt sich dann

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}\alpha} g(\alpha) &= \frac{g_z'(\alpha)g_n(\alpha) - g_z(\alpha)g_n'(\alpha)}{g_n^2(\alpha)} = \frac{\frac{g_z(\alpha)g_n(\alpha)}{\gamma\sqrt{(\gamma^{-1}(\alpha-1)+1)^2-1}} - \frac{g_z(\alpha)g_n(\alpha)}{\sqrt{\alpha^2-1}}}{g_n^2(\alpha)} \\ &= \frac{g_z(\alpha)g_n(\alpha)}{g_n^2(\alpha)} \left(\frac{1}{\gamma\sqrt{(\gamma^{-1}(\alpha-1)+1)^2-1}} - \frac{1}{\sqrt{\alpha^2-1}} \right).\end{aligned}$$

Da sowohl g_z als auch g_n nach $\mathbb{R}_{\geq 1}$ abbilden, reicht es, den Term in der Klammer näher zu untersuchen. Die Behauptung $g'(\alpha) > 0$ gilt genau dann, wenn

$$\sqrt{\alpha^2 - 1} > \gamma\sqrt{(\gamma^{-1}(\alpha-1)+1)^2 - 1}$$

erfüllt und damit die Differenz ebenfalls positiv ist. Dies ist äquivalent zu

$$\alpha^2 - 1 > \gamma^2 \left(\gamma^{-1}(\alpha-1)+1\right)^2 - \gamma^2.$$

Für die rechte Seite und $\alpha \in \mathbb{R}_{>1}$ gilt

$$\begin{aligned}\gamma^2 \left(\gamma^{-1}(\alpha-1)+1\right)^2 - \gamma^2 &= (\alpha-1)^2 + 2\gamma(\alpha-1) \\ &< \alpha^2 - 2\alpha + 1 + 2(\alpha-1) \\ &= \alpha^2 - 1,\end{aligned}$$

die Funktion ϑ ist damit für $\alpha \in \mathbb{R}_{>1}$ streng monoton wachsend.
Die Behauptung $\vartheta(1) = 1$ ergibt sich direkt durch

$$\vartheta(1) = \frac{\varpi^\dagger\left(\frac{\varpi(1)-1}{\gamma}+1\right)}{1} = \varpi^\dagger\left(\frac{1-1+\gamma}{\gamma}\right) = \varpi^\dagger(1) = 1.$$

Für den Grenzwert im Unendlichen betrachte zunächst ein paar simplere Grenzwerte, um die gewünschte Aussage dann auf diese zurückzuführen. Für die Funktion ϖ gilt offensichtlich mit den Regeln für Grenzwerte

$$\lim_{\alpha\to\infty} \frac{\varpi(\alpha)}{\alpha} = \lim_{\alpha\to\infty} \frac{\alpha+\frac{1}{\alpha}}{2\alpha} = \frac{1}{2} \lim_{\alpha\to\infty} \left(1 + \frac{1}{\alpha^2}\right) = \frac{1}{2}.$$

Für die Rechtsinverse $\varpi^\dagger$ auf $\mathbb{R}_{\geq 1}$ gilt mit (4.1.5)

$$\lim_{\alpha\to\infty} \frac{\varpi^\dagger(\alpha)}{\alpha} = \lim_{\alpha\to\infty} \frac{\alpha+\sqrt{\alpha^2-1}}{\alpha} = 1 + \lim_{\alpha\to\infty} \frac{\sqrt{\alpha^2-1}}{\alpha} = 2$$

und der Term $\frac{\varpi(\alpha)-1}{\gamma} + 1$ strebt aufgrund des monotonen Wachstums von ϖ (4.1.4) für

$\alpha \to \infty$ ebenfalls gegen Unendlich. Mit diesen Grenzwerten ergibt sich

$$\begin{aligned}
\lim_{\alpha\to\infty} \vartheta(\alpha) &= \lim_{\alpha\to\infty} \frac{\varpi^\dagger\left(\frac{\varpi(\alpha)-1}{\gamma}+1\right)}{\alpha} \\
&= \lim_{\alpha\to\infty} \frac{\varpi^\dagger\left(\frac{\varpi(\alpha)-1}{\gamma}+1\right)}{\frac{\varpi(\alpha)-1}{\gamma}+1} \frac{\frac{\varpi(\alpha)-1}{\gamma}+1}{\alpha} \\
&= \lim_{\alpha\to\infty} \frac{\varpi^\dagger\left(\frac{\varpi(\alpha)-1}{\gamma}+1\right)}{\frac{\varpi(\alpha)-1}{\gamma}+1} \lim_{\alpha\to\infty} \frac{\frac{\varpi(\alpha)-1}{\gamma}+1}{\alpha} \\
&= 2 \lim_{\alpha\to\infty} \frac{\varpi(\alpha)-1+\gamma}{\gamma\alpha} = 2\frac{1}{\gamma}\left(\lim_{\alpha\to\infty} \frac{\varpi(\alpha)}{\alpha} - \lim_{\alpha\to\infty} \frac{1-\gamma}{\alpha}\right) \\
&= 2\frac{1}{\gamma}\frac{1}{2} = \frac{1}{\gamma}.
\end{aligned}$$

□

Lemma 4.23 (Geschachtelte Bernstein-Ellipsen)
Seien ein $L \in \underline{p_{\mathcal{I}}}$, $\varrho \in \mathbb{R}_{>1}$ und $\sigma := \vartheta(\varrho)$ gegeben. Dann gilt $\sigma > 1$. Sei weiter ein $i \in \underline{6}$ gegeben. Wenn für alle $\ell \in \underline{L}$, die dazugehörigen Intervalle $I_{\ell,i}$ die Kontraktionseigenschaft 4.3.5 erfüllen, weisen auch die auf die entsprechenden Intervalle transformierten Bernstein-Ellipsen eine Schachtelungseigenschaft auf, es gilt für $I_\ell = I_{\ell,i}$

$$^{I_\ell}D_{\sigma^k\varrho} \subset {}^{I_{\ell-k}}D_\varrho \qquad \text{für alle } \ell \in \underline{L}, k \in \underline{\ell}.$$

Beweis: Dass $\sigma > 1$ gilt, folgt direkt mit Lemma 4.22, da $\varrho > 1$ erfüllt und ϑ streng monoton wachsend ist. Für die Schachtelung der Bernstein-Ellipsen betrachte zunächst für ein $\ell \in \underline{L}$ die Intervalle I_ℓ und $I_{\ell-1}$ und nutze eine verkürzte Schreibweise der Intervallgrenzen mit $I_\ell = [a_\ell, b_\ell]$ sowie $I_{\ell-1} = [a_{\ell-1}, b_{\ell-1}]$. Verwende weiter die Abbildung $\Phi_{I_{\ell-1}}$, um $\Phi_{I_{\ell-1}}(I_\ell) \subset [-1, 1]$ zu erhalten. Setze

$$\bar{a}_\ell := \frac{2}{b_{\ell-1}-a_{\ell-1}}\left(a_\ell - \frac{b_{\ell-1}+a_{\ell-1}}{2}\right) \in \mathbb{R},$$

dann gilt $\Phi_{I_{\ell-1}}(\bar{a}_\ell) = a_\ell$, denn

$$\begin{aligned}
\Phi_{I_{\ell-1}}(\bar{a}_\ell) &= \frac{b_{\ell-1}+a_{\ell-1}}{2} + \frac{b_{\ell-1}-a_{\ell-1}}{2}\bar{a}_\ell \\
&= \frac{b_{\ell-1}+a_{\ell-1}}{2} + \frac{b_{\ell-1}-a_{\ell-1}}{2}\left(\frac{2}{b_{\ell-1}-a_{\ell-1}}\left(a_\ell - \frac{b_{\ell-1}+a_{\ell-1}}{2}\right)\right) \\
&= \frac{b_{\ell-1}+a_{\ell-1}}{2} + a_\ell - \frac{b_{\ell-1}+a_{\ell-1}}{2} = a_\ell.
\end{aligned}$$

Analog definiere die rechte Intervallgrenze

$$\bar{b}_\ell := \frac{2}{b_{\ell-1}-a_{\ell-1}}\left(b_\ell - \frac{b_{\ell-1}+a_{\ell-1}}{2}\right) \in \mathbb{R},$$

für die dann $\Phi_{I_{\ell-1}}(\bar{b}_\ell) = b_\ell$ gilt. Weiterhin gilt $\bar{a}_\ell \geq -1$, denn mit $a_{\ell-1} \leq a_\ell$ folgt

$$\begin{aligned}\bar{a}_\ell &= \tfrac{2}{b_{\ell-1}-a_{\ell-1}} \left(a_\ell - \tfrac{b_{\ell-1}+a_{\ell-1}}{2}\right) \\ &\geq \tfrac{2}{b_{\ell-1}-a_{\ell-1}} \left(a_{\ell-1} - \tfrac{b_{\ell-1}+a_{\ell-1}}{2}\right) \\ &= \tfrac{2}{b_{\ell-1}-a_{\ell-1}} \left((-1)\tfrac{b_{\ell-1}-a_{\ell-1}}{2}\right) = -1.\end{aligned}$$

Ebenso folgt $\bar{b}_\ell \leq 1$ aus $b_{\ell-1} \geq b_\ell$ mit

$$\begin{aligned}\bar{b}_\ell &= \tfrac{2}{b_{\ell-1}-a_{\ell-1}} \left(b_\ell - \tfrac{b_{\ell-1}+a_{\ell-1}}{2}\right) \\ &\leq \tfrac{2}{b_{\ell-1}-a_{\ell-1}} \left(b_{\ell-1} - \tfrac{b_{\ell-1}+a_{\ell-1}}{2}\right) \\ &= \tfrac{2}{b_{\ell-1}-a_{\ell-1}} \left(\tfrac{b_{\ell-1}-a_{\ell-1}}{2}\right) = 1.\end{aligned}$$

Aus der Kontraktionseigenschaft der betrachteten Intervalle $I_\ell, I_{\ell-1}$ ergibt sich, dass $\gamma \geq 2^{-1}(\bar{b}_\ell - \bar{a}_\ell)$ gilt. Zusammen mit dem monoton fallend Verhalten von ϑ für wachsende γ in $(0,1)$ folgt

$$\sigma\varrho = \vartheta(\varrho)\varrho \leq \frac{\varpi^\dagger\left(\frac{\varpi(\varrho)-1}{\gamma}+1\right)}{\varrho}\varrho = \varpi^\dagger\left(\tfrac{\varpi(\varrho)-1}{\gamma}+1\right).$$

Entsprechend liefert Lemma 4.21, dass dann ${}^{\bar{a}_\ell,\bar{b}_\ell}D_{\sigma\varrho} \subset D_\varrho$ gilt und damit ergibt sich

$${}^{I_\ell}D_{\sigma\varrho} = \Phi_{I_{\ell-1}}({}^{\bar{a}_\ell,\bar{b}_\ell}D_{\sigma\varrho}) \subset \Phi_{I_{\ell-1}}(D_\varrho) = {}^{I_{\ell-1}}D_\varrho.$$

Da die Funktion ϑ nach Lemma 4.22 streng monoton wachsend ist, folgt für ein $\widehat{\varrho} \geq \varrho$

$$\sigma\widehat{\varrho} = \vartheta(\varrho)\widehat{\varrho} \leq \vartheta(\widehat{\varrho})\widehat{\varrho}$$

und damit ergibt sich dann auch

$${}^{I_\ell}D_{\sigma\widehat{\varrho}} \subset {}^{I_{\ell-1}}D_{\widehat{\varrho}} \qquad \text{für alle } \widehat{\varrho} \geq \varrho. \tag{4.3.7}$$

Der letzte Teil kann für $\ell \in \underline{L_\rfloor}$ mit einer einfachen Induktion über k gezeigt werden.
I.A. Sei $k = 1$, dann gilt, wie oben schon gezeigt, ${}^{I_\ell}D_{\sigma\varrho} \subset {}^{I_{\ell-1}}D_\varrho$.
I.V. Sei nun ein $k \in \underline{\ell-1_\rfloor}$ so gegeben, dass die Behauptung gilt.
I.S. Betrachte $k+1$ und setze $\widehat{\sigma} := \sigma^k\varrho$. Dann gilt $\widehat{\sigma} > \varrho$ sowie $\sigma\widehat{\sigma} = \sigma\sigma^k\varrho = \sigma^{k+1}\varrho$, so dass mit der Induktionsvoraussetzung

$${}^{I_\ell}D_{\sigma^{k+1}\varrho} = {}^{I_\ell}D_{\sigma\widehat{\sigma}} \overset{(4.3.7)}{\subset} {}^{I_{\ell-1}}D_{\widehat{\sigma}} = {}^{I_{\ell-1}}D_{\sigma^k\varrho} \overset{I.V.}{\subset} {}^{I_{\ell-(k+1)}}D_\varrho$$

folgt. □

Um die Beweise der gewünschten Fehler- und Stabilitätsaussage nicht zu lang werden zu lassen, werden noch zwei weitere notwendige Abschätzungen vorweg bewiesen.
Dazu soll zunächst eine Aussage zum Fehler der Interpolation ohne ebene Welle auf geschachtelten Bernstein-Ellipsen gemacht werden. Dieser Ansatz stammt ebenfalls aus einer unveröffentlichten Arbeit zur iterierten Interpolation von Herrn Börm [9].

Lemma 4.24 (Interpolationsfehler auf geschachtelten Bernstein-Ellipsen)
Seien ein $L \in \underline{p_{\mathcal{I}}}$, ein $i \in \underline{6}$ sowie ein $\varrho \in \mathbb{R}_{>1}$ und $\sigma := \vartheta(\varrho)$ gegeben. Für alle $\ell \in \underline{L}$ erfüllen die dazugehörigen Intervalle die Kontraktionseigenschaft 4.3.5. Weiter sei der betrachtete Interpolationsoperator von Ordnung $m \in \mathbb{N}_0$ stabil nach Bemerkung 6. Zu jedem $q \in (\sigma^{-1}, 1)$ existiert dann ein $c_{q,\varrho}$ mit

$$c_{q,\varrho} := \max\left\{(1+\Lambda(m+1)^{\lambda})(\sigma q)^{-m}\tfrac{2}{\sigma-1} \mid m \in \mathbb{N}_0\right\}.$$

Für alle $\ell \in \underline{L}$ sei $I_\ell = I_{\ell,i}$, dann gilt für alle auf der transformierten Bernstein-Ellipse ${}^{I_{\ell-1}}D_\varrho$ holomorphen Funktionen f für den Interpolationsfehler

$$\|f - \mathfrak{I}_{I_\ell}[f]\|_{\infty,{}^{I_\ell}D_\varrho} \leq c_{q,\varrho} q^m \|f\|_{\infty,{}^{I_{\ell-1}}D_\varrho}$$

sowie für den Interpolationsoperator

$$\|\mathfrak{I}_{I_\ell}[f]\|_{\infty,{}^{I_\ell}D_\varrho} \leq (1 + c_{q,\varrho} q^m)\|f\|_{\infty,{}^{I_{\ell-1}}D_\varrho}.$$

Beweis: Vorab sei bemerkt, dass die Wahl von q sicherstellt, dass $\sigma q > 1$ erfüllt. Entsprechend ist $c_{q,\varrho}$ als das Maximum über ein Produkt eines polynomiell in m wachsenden Terms $(1+\Lambda(m+1)^{\lambda})$, einer Konstante $\frac{2}{\sigma-1}$ und eines in m exponentiell fallenden Terms $(\sigma q)^{-m}$ wohldefiniert.
Seien ein $\ell \in \underline{L}$ und eine auf ${}^{I_{\ell-1}}D_\varrho$ holomorphe Funktion f gegeben. Schreibe die Interpolation zunächst mit Hilfe der Abbildung Φ_{I_ℓ} auf kompliziertere Weise. Es gilt

$$\begin{aligned}\|f - \mathfrak{I}_{I_\ell}[f]\|_{\infty,{}^{I_\ell}D_\varrho} &= \|(f\circ\Phi_{I_\ell})\circ\Phi_{I_\ell}^{-1} - (\mathfrak{I}_{[-1,1]}[f\circ\Phi_{I_\ell}])\circ\Phi_{I_\ell}^{-1}\|_{\infty,\Phi_{I_\ell}(D_\varrho)}\\ &= \|f\circ\Phi_{I_\ell} - \mathfrak{I}_{[-1,1]}[f\circ\Phi_{I_\ell}]\|_{\infty,D_\varrho}.\end{aligned}$$

Auf diese Formulierung kann das Lemma 4.9 mit $D_{\sigma\varrho}$ als Bernstein-Ellipse auf der rechten Seite angewendet werden

$$\begin{aligned}\|f\circ\Phi_{I_\ell} - \mathfrak{I}_{[-1,1]}[f\circ\Phi_{I_\ell}]\|_{\infty,D_\varrho} &\leq (1+\Lambda_m)\left(\tfrac{\varrho}{\sigma\varrho}\right)^m \tfrac{2}{\sigma\varrho(\varrho)^{-1}-1}\|f\circ\Phi_{I_\ell}\|_{\infty,D_{\sigma\varrho}}\\ &= (1+\Lambda_m)\,(\sigma)^{-m}\,\tfrac{2}{\sigma-1}\|f\circ\Phi_{I_\ell}\|_{\infty,D_{\sigma\varrho}}\\ &= (1+\Lambda_m)\sigma^{-m}\tfrac{2}{\sigma-1}\|f\|_{\infty,{}^{I_\ell}D_{\sigma\varrho}}.\end{aligned}$$

Mit dem Lemma 4.23 zu den geschachtelten Bernstein-Ellipsen folgt dann

$$(1+\Lambda_m)\sigma^{-m}\tfrac{2}{\sigma-1}\|f\|_{\infty,{}^{I_\ell}D_{\sigma\varrho}} \leq (1+\Lambda_m)\sigma^{-m}\tfrac{2}{\sigma-1}\|f\|_{\infty,{}^{I_{\ell-1}}D_\varrho}.$$

Nach den Voraussetzungen existieren Konstanten $\Lambda, \lambda \in \mathbb{R}_{\geq 1}$ mit

$$\Lambda_m \leq \Lambda(m+1)^\lambda \qquad \text{für alle } m \in \mathbb{N}_0,$$

so dass mit der Konstanten $c_{q,\varrho}$ der erste Teil der Behauptung folgt

$$\begin{aligned}(1+\Lambda_m)\sigma^{-m}\tfrac{2}{\sigma-1} &\leq (1+\Lambda(m+1)^\lambda)(\sigma q)^{-m} q^m \tfrac{2}{\sigma-1} \\ &\leq c_{q,\varrho} q^m.\end{aligned}$$

Der zweite Teil folgt aus dem ersten zusammen mit dem Lemma 4.23

$$\begin{aligned}\|\mathfrak{I}_{I_\ell}[f]\|_{\infty,\,^{I_\ell}D_\varrho} &= \|f - f + \mathfrak{I}_{I_\ell}[f]\|_{\infty,\,^{I_\ell}D_\varrho} \\ &\overset{\Delta}{\leq} \|f\|_{\infty,\,^{I_\ell}D_\varrho} + \|-f+\mathfrak{I}_{I_\ell}[f]\|_{\infty,\,^{I_\ell}D_\varrho} \\ &\leq \|f\|_{\infty,\,^{I_\ell}D_\varrho} + c_{q,\varrho} q^m \|f\|_{\infty,\,^{I_{\ell-1}}D_\varrho} \\ &\leq \|f\|_{\infty,\,^{I_\ell}D_{\sigma\varrho}} + c_{q,\varrho} q^m \|f\|_{\infty,\,^{I_{\ell-1}}D_\varrho} \\ &\leq (1 + c_{q,\varrho} q^m)\|f\|_{\infty,\,^{I_{\ell-1}}D_\varrho}.\end{aligned}$$

□

Die Argumentation im Lemma 4.24 kann auch genutzt werden, um den Interpolationsfehler gegen eine Funktionsauswertung auf einer auf das Ursprungsintervall I_ℓ transformierten Bernstein-Ellipse mit einer größeren Halbachsensumme abzuschätzen [9].

Bemerkung 23 (Größere Bernstein-Ellipsen): *Seien ein $L \in \underline{p_{\mathcal{I}}}$, ein $i \in \underline{6}$ sowie $\varrho \in \mathbb{R}_{\geq 1}$, $\tau \in \mathbb{R}_{\geq 1}$ und $\sigma > 1$ gegeben. Weiter sei der betrachtete Interpolationsoperator von Ordnung $m \in \mathbb{N}_0$ stabil nach Bemerkung 6. Dann existiert zu jedem $q \in (\sigma^{-1}, 1)$ ein $c_{q,\varrho}$ wie in Lemma 4.24. Für alle $\ell \in \underline{L}$ sei $I_\ell = I_{\ell,i}$. Dann gilt für alle auf der transformierten Bernstein-Ellipse $^{I_{\ell-1}}D_{\sigma\tau\varrho}$ holomorphen Funktionen f für den Interpolationsfehler*

$$\|f - \mathfrak{I}_{I_\ell}[f]\|_{\infty,\,^{I_\ell}D_\varrho} \leq c_{q,\varrho} q^m \tau^{-m} \|f\|_{\infty,\,^{I_\ell}D_{\sigma\tau\varrho}}.$$

Beweis: Zunächst wird auch hier die Interpolation mit Hilfe der Abbildung Φ_{I_ℓ} auf kompliziertere Weise geschrieben

$$\|f - \mathfrak{I}_{I_\ell}[f]\|_{\infty,\,^{I_\ell}D_\varrho} = \|f \circ \Phi_{I_\ell} - \mathfrak{I}_{[-1,1]}[f \circ \Phi_{I_\ell}]\|_{\infty, D_\varrho}.$$

Da $\sigma > 1$, gilt $\sigma\tau\varrho > \varrho$, so dass auf diese Formulierung das Lemma 4.9 mit $D_{\sigma\tau\varrho}$ als Bernstein-Ellipse auf der rechten Seite angewendet werden kann. Die Behauptung folgt dann durch Abschätzen mit $c_{q,\varrho}$

$$\begin{aligned}\|f \circ \Phi_{I_\ell} - \mathfrak{I}_{[-1,1]}[f \circ \Phi_{I_\ell}]\|_{\infty, D_\varrho} &\leq (1+\Lambda_m)\left(\tfrac{\varrho}{\sigma\tau\varrho}\right)^m \tfrac{2}{\sigma\tau\varrho(\varrho)^{-1}-1} \|f \circ \Phi_{I_\ell}\|_{\infty, D_{\sigma\tau\varrho}} \\ &\leq (1+\Lambda_m)\,(\sigma)^{-m}\,\tfrac{2}{\sigma-1}\,(\tau)^{-m}\,\|f \circ \Phi_{I_\ell}\|_{\infty, D_{\sigma\tau\varrho}} \\ &\leq c_{q,\varrho} q^m \tau^{-m} \|f\|_{\infty,\,^{I_\ell}D_{\sigma\tau\varrho}}.\end{aligned}$$

□

Um dieses Ergebnis auch auf die Interpolation mit der zusätzlichen ebenen Welle übertragen zu können, ist es notwendig, das Maximum einer ebenen Welle mit Wellenzahl $\kappa \in \mathbb{R}_{\geq 0}$ auf einer transformierten Bernstein-Ellipse zu bestimmen. Nach Bemerkung 20 nimmt die ebene Welle als eine auf ganz $\mathbb{C}$ holomorphe Funktion ihr Maximum auf dem Rand der Bernstein-Ellipse an. Wenn ${}^{a,b}D_\alpha$ die zugrundeliegende auf $[a,b]$ transformierte Bernstein-Ellipse und ein beliebiges, aber festes $\widehat{c} \in [-2,2]$ gegeben sind, besteht das dazugehörige Problem darin, $z \in \partial({}^{a,b}D_\alpha)$ so zu finden, dass

$$\left|e^{\mathrm{i}\kappa\widehat{c}z}\right| \geq \left|e^{\mathrm{i}\kappa\widehat{c}w}\right| \qquad \text{für alle } w \in {}^{a,b}\overline{D}_\alpha$$

gilt. Falls $\widehat{c} = 0$ sein sollte, folgt sofort $e^{\mathrm{i}\kappa 0 z} = 1$. Für den Fall $\widehat{c} \neq 0$ nutze, dass ${}^{a,b}D_\alpha = \Phi_{[a,b]}(D_\alpha)$ gilt, und suche stattdessen ein $z \in \partial D_\alpha$ mit

$$\left|e^{\mathrm{i}\kappa\widehat{c}\Phi_{[a,b]}(z)}\right| \geq \left|e^{\mathrm{i}\kappa\widehat{c}\Phi_{[a,b]}(w)}\right| \qquad \text{für alle } w \in \overline{D}_\alpha.$$

Setze zunächst die Abbildungsvorschrift von $\Phi_{[a,b]}$ ein

$$\left|e^{\mathrm{i}\kappa\widehat{c}\Phi_{[a,b]}(z)}\right| = \left|e^{\mathrm{i}\kappa\widehat{c}\left(\frac{b+a}{2}+\frac{b-a}{2}z\right)}\right| = \left|e^{\mathrm{i}\kappa\widehat{c}\frac{b+a}{2}}\right| \left|e^{\mathrm{i}\kappa\widehat{c}\frac{b-a}{2}z}\right| = \left|e^{\mathrm{i}\kappa\widehat{c}\frac{b-a}{2}z}\right|.$$

Da z eine komplexe Zahl ist, kann die Identität $z = d + \mathrm{i}g$ für geeignete $d, g \in \mathbb{R}$ genutzt werden, was zu

$$\left|e^{\mathrm{i}\kappa\widehat{c}\frac{b-a}{2}z}\right| = \left|e^{\mathrm{i}\kappa\widehat{c}\frac{b-a}{2}(d+\mathrm{i}g)}\right| = \left|e^{\mathrm{i}\kappa\widehat{c}\frac{b-a}{2}d}\right| \left|e^{-\kappa\widehat{c}\frac{b-a}{2}g}\right| = \left|e^{-\kappa\widehat{c}\frac{b-a}{2}g}\right|$$

führt. Nutze schließlich die Monotonie der Exponentialfunktion

$$\left|e^{-\kappa\widehat{c}\frac{b-a}{2}g}\right| \leq e^{\left|\kappa\widehat{c}\frac{b-a}{2}g\right|} \leq e^{\kappa|\widehat{c}|\left|\frac{b-a}{2}\right||g|}.$$

Da für $z \in D_\alpha$ auch $\bar{z} \in D_\alpha$ gilt, reicht es, $g > 0$ zu untersuchen. Jeder Punkt z auf dem Rand der Bernstein-Ellipse ∂D_α erfüllt

$$\frac{d^2}{a_\alpha^2} + \frac{g^2}{b_\alpha^2} = 1,$$

und da nach der Vorbetrachtung der Realteil d von z irrelevant ist, reicht es aus den Imaginärteil zu maximieren, also $g = b_\alpha$ zu wählen. Entsprechend gilt

$$\left|e^{\mathrm{i}\kappa\widehat{c}w}\right| \leq e^{\kappa|\widehat{c}|\left|\frac{b-a}{2}\right|b_\alpha} \qquad \text{für alle } w \in {}^{a,b}\overline{D}_\alpha. \tag{4.3.8}$$

Gesucht wird für $i \in \underline{6}_\rfloor$, $\ell \in \underline{L}_\rfloor$ und $\widehat{c} = \dot{c}^i_{\ell-1} - \dot{c}^i_\ell$

$$\max\left\{|e^{\mathrm{i}\kappa\widehat{c}z}| \mid z \in \partial({}^{I_{\ell-1,i}}D_\varrho)\right\}.$$

Um das Maximum unabhängig vom konkreten κ, $i \in \underline{6}$ und dem Intervall $I_{\ell-1,i}$ abschätzen zu können, nehme an, dass zu jedem $\varrho \in \mathbb{R}_{>1}$ ein $\overline{\gamma}_\varrho > 0$ existiert mit

$$\kappa|\dot{c}^i_{\ell-1} - \dot{c}^i_\ell|\tfrac{|I_{\ell-1,i}|}{2} b_\varrho \leq \overline{\gamma}_\varrho \qquad \text{für alle } i \in \underline{6}, \ell \in \underline{L}. \tag{4.3.9}$$

Der Term $\kappa|\dot{c}^i_{\ell-1} - \dot{c}^i_\ell|\frac{|I_{\ell-1,i}|}{2}$ wird unabhängig von $i \in \underline{6}$ und $\ell \in \underline{L}$ durch die erste Zulässigkeitsbedingung (2.3.2a) beschränkt, so dass $\overline{\gamma}_\varrho$ tatsächlich nur von der Zulässigkeitsbedingung und ϱ abhängig ist. Auch Terme, bei denen die Differenz von nicht direkt aufeinander folgenden Richtungen betrachtet wird, lassen sich mit dieser Annahme beschränken.

Lemma 4.25
Seien für ein $L \in \underline{p_{\mathcal{I}}}$ und für alle $i \in \underline{6}$ Intervallsequenzen, die die Kontraktionseigenschaft 4.3.5 für ein $\mathcal{C}_k \in (0,1)$ erfüllen, und eine dazugehörige Sequenz an Richtungen gegeben. Weiter sei ein $\varrho \in \mathbb{R}_{>1}$ gegeben und es existiere ein $\overline{\gamma}_\varrho > 0$, so dass die Bedingung (4.3.9) *erfüllt ist. Dann gilt*

$$\kappa|\dot{c}^i_\ell - \dot{c}^i_0|\tfrac{|I_{\ell-1,i}|}{2} b_\varrho \leq \overline{\gamma}_\varrho \sum_{j=0}^{\ell-1} \mathcal{C}_k^j \qquad \textit{für alle } \ell \in \underline{L}, i \in \underline{6}.$$

Beweis: Sei $i \in \underline{6}$ gegeben. Zeige die Behauptung per Induktion über ℓ.
I.A. Sei $\ell = 1$, dann gilt direkt mit der Bedingung (4.3.9)

$$\kappa|\dot{c}^i_1 - \dot{c}^i_0|\tfrac{|I_{0,i}|}{2} b_\varrho \leq \overline{\gamma}_\varrho = \overline{\gamma}_\varrho \sum_{j=0}^{\ell-1} \mathcal{C}_k^j.$$

I.V. Sie ein $\ell \in \underline{L-1}$ so gegeben, dass die Behauptung gilt.
I.S. Betrachte $\ell + 1$. Es folgt

$$\begin{aligned}\kappa|\dot{c}^i_{\ell+1} - \dot{c}^i_0|\tfrac{|I_{\ell,i}|}{2} b_\varrho &= \kappa|\dot{c}^i_{\ell+1} - \dot{c}^i_\ell + \dot{c}^i_\ell - \dot{c}^i_0|\tfrac{|I_{\ell,i}|}{2} b_\varrho \\ &\overset{\Delta}{\leq} \kappa|\dot{c}^i_{\ell+1} - \dot{c}^i_\ell|\tfrac{|I_{\ell,i}|}{2} b_\varrho + \kappa|\dot{c}^i_\ell - \dot{c}^i_0|\tfrac{|I_{\ell,i}|}{2} b_\varrho.\end{aligned}$$

Der erste Term kann direkt mit der Bedingung (4.3.9) beschränkt werden, für den zweiten Term nutze die Kontraktionseigenschaft der Intervalle und die Induktionsvoraussetzung

$$\begin{aligned}\kappa|\dot{c}^i_{\ell+1} - \dot{c}^i_\ell|\tfrac{|I_{\ell,i}|}{2} b_\varrho + \kappa|\dot{c}^i_\ell - \dot{c}^i_0|\tfrac{|I_\ell|}{2} b_\varrho &\leq \overline{\gamma}_\varrho + \kappa|\dot{c}^i_\ell - \dot{c}^i_0|\tfrac{|I_{\ell,i}|}{2}\tfrac{|I_{\ell-1,i}|}{|I_{\ell-1,i}|} b_\varrho \\ &= \overline{\gamma}_\varrho + \kappa|\dot{c}^i_\ell - \dot{c}^i_0|\tfrac{|I_{\ell-1,i}|}{2} b_\varrho \mathcal{C}_k \\ &\overset{\text{I.V.}}{\leq} \overline{\gamma}_\varrho + \overline{\gamma}_\varrho \mathcal{C}_k \sum_{j=0}^{\ell-1} \mathcal{C}_k^j \\ &= \overline{\gamma}_\varrho \left(1 + \sum_{j=1}^{\ell} \mathcal{C}_k^j\right) = \overline{\gamma}_\varrho \sum_{j=0}^{\ell} \mathcal{C}_k^j.\end{aligned}$$

□

Bemerkung 24 : *Die Abhängigkeit der Schranke aus Lemma 4.25 von dem konkreten ℓ kann eliminiert werden. Da $\mathcal{C}_k \in (0,1)$, konvergiert die geometrische Reihe und es folgt*

$$\kappa|\dot{c}_\ell^i - \dot{c}_0^i|\tfrac{|I_{\ell-1,i}|}{2}b_\varrho \leq \overline{\gamma}_\varrho \sum_{j=0}^{\ell-1}\mathcal{C}_k^j \leq \overline{\gamma}_\varrho(1-\mathcal{C}_k)^{-1}.$$

Setze $\gamma_\varrho := \overline{\gamma}_\varrho(1-\mathcal{C}_k)^{-1} > 0$, damit ergibt sich

$$\kappa|\dot{c}_\ell^i - \dot{c}_0^i|\tfrac{|I_{\ell-1,i}|}{2}b_\varrho \leq \gamma_\varrho \qquad \textit{für alle } \ell \in \underline{L\rfloor}, i \in \underline{6\rfloor}. \tag{4.3.10}$$

Natürlich gilt die Abschätzung auch, wenn eine andere Richtung $\dot{c}_k^i$ für $k \in \underline{\ell\rfloor}$ statt $\dot{c}_0^i$ betrachtet wird.

Mit dieser Schranke ergibt sich dann für $\widehat{c} = \dot{c}_\ell^i - \dot{c}_0^i$

$$\max\left\{|e^{\mathrm{i}\kappa\widehat{c}z}| \mid z \in {}^{I_{\ell-1,i}}D_\varrho\right\} \leq e^{\gamma_\varrho} \qquad \text{für alle } \ell \in \underline{L\rfloor}, i \in \underline{6\rfloor}. \tag{4.3.11}$$

Da die beteiligten Intervalle die Kontraktionseigenschaft erfüllen, gilt $|I_{\ell-1,i}| > |I_{\ell,i}|$ und damit folgt dann auch

$$\kappa|\dot{c}_\ell^i - \dot{c}_0^i|\tfrac{|I_{\ell,i}|}{2}b_\varrho \leq \kappa|\dot{c}_\ell^i - \dot{c}_0^i|\tfrac{|I_{\ell-1,i}|}{2}b_\varrho \leq \gamma_\varrho,$$

so dass auch

$$\max\left\{|e^{\mathrm{i}\kappa\widehat{c}z}| \mid z \in {}^{I_{\ell,i}}D_\varrho\right\} \leq e^{\gamma_\varrho} \qquad \text{für alle } \ell \in \underline{L\rfloor}, i \in \underline{6\rfloor}$$

erfüllt ist. Damit ist alles beisammen, um die Reinterpolation mit ebenen Wellen genauer zu untersuchen. Zunächst soll dabei nur ein Schritt der Reinterpolation betrachtet und damit eine Aussage zum Fehler

$$\|e^{\mathrm{i}\kappa\dot{c}_{\ell-1}^i\cdot}p - \mathfrak{I}_{I_\ell}^{i,c_\ell}[e^{\mathrm{i}\kappa\dot{c}_{\ell-1}^i\cdot}p]\|_{\infty,I_\ell} \qquad \text{für alle } i \in \underline{6\rfloor}, \ell \in \underline{L\rfloor}, p \in \Pi_m,$$

wobei $e^{\mathrm{i}\kappa\dot{c}_{\ell-1}^i\cdot}$ die ebene Welle mit einem beliebigen Argument sein soll, getroffen werden. Die Analyse dieses Fehlers wird allgemeiner für eine holomorphe Funktion f durchgeführt, welche die Rolle der speziellen holomorphen Funktion $e^{\mathrm{i}\kappa\dot{c}_{\ell-1}^i\cdot}p$ übernimmt.

Lemma 4.26 (Einmalige Reinterpolation)
Seien für $L \in \underline{p_{\mathcal{I}}\rfloor}$ und ein $i \in \underline{6\rfloor}$ eine Intervallsequenz, welche die Kontraktionseigenschaft (4.3.5) *erfüllt, eine Sequenz an Richtungen sowie $\varrho \in \mathbb{R}_{>1}$ und $\sigma := \vartheta(\varrho)$ gegeben und es existiere ein $\gamma_\varrho \in \mathbb{R}_{>0}$, das* (4.3.10) *erfüllt. Weiter sei der betrachtete Interpolationsoperator von Ordnung $m \in \mathbb{N}_0$ stabil nach Bemerkung 6. Dann gibt es zu jedem $q \in (\sigma^{-1},1)$ ein $c_{q,\varrho}$ wie in Lemma 4.24, mit dem für alle $\ell \in \underline{L\rfloor}$ und für alle auf ${}^{I_{\ell-1}}D_\varrho$ mit $I_{\ell-1} = I_{\ell-1,i}$ holomorphen Funktionen f*

$$\|f - \mathfrak{I}_{I_\ell}^{i,c_\ell}[f]\|_{\infty,I_\ell} \leq e^{\gamma_\varrho}c_{q,\varrho}q^m\|e^{-\mathrm{i}\kappa\dot{c}_0^i\cdot}f\|_{\infty,{}^{I_{\ell-1}}D_\varrho}$$

sowie

$$\|e^{-\mathrm{i}\kappa\dot{c}_0^i\cdot}\mathfrak{I}_{I_\ell}^{i,c_\ell}[f]\|_{\infty,{}^{I_\ell}D_\varrho} \leq (1+e^{2\gamma_\varrho}c_{q,\varrho}q^m)\|e^{-\mathrm{i}\kappa\dot{c}_0^i\cdot}f\|_{\infty,{}^{I_{\ell-1}}D_\varrho}$$

gelten.

Beweis: Seien ein $\ell \in \underline{L}_\rfloor$ und eine auf ${}^{I_{\ell-1}}D_\varrho$ holomorphe Funktion f gegeben. Führe die Interpolation mit ebener Welle zunächst auf die normale Interpolation auf I_ℓ zurück und nutze die Funktion $v := e^{-\mathrm{i}\kappa\dot{c}_\ell^i\cdot}f$, um die Schritte übersichtlich zu halten

$$\begin{aligned}\|f-\mathfrak{I}_{I_\ell}^{i,c_\ell}[f]\|_{\infty,I_\ell} &= \|e^{\mathrm{i}\kappa\dot{c}_\ell^i\cdot}\left(e^{-\mathrm{i}\kappa\dot{c}_\ell^i\cdot}f-\mathfrak{I}_{I_\ell}[e^{-\mathrm{i}\kappa\dot{c}_\ell^i\cdot}f]\right)\|_{\infty,I_\ell}\\ &\overset{(4.2.1)}{=} \|v-\mathfrak{I}_{I_\ell}[v]\|_{\infty,I_\ell}.\end{aligned}$$

Grundsätzlich ist $I_\ell \subset {}^{I_\ell}D_\varrho$, entsprechend gilt

$$\|v-\mathfrak{I}_{I_\ell}[v]\|_{\infty,I_\ell} \leq \|v-\mathfrak{I}_{I_\ell}[v]\|_{\infty,{}^{I_\ell}D_\varrho}.$$

Da die Funktion v auf der Bernstein-Ellipse ${}^{I_{\ell-1}}D_\varrho$ als Produkt der holomorphen Funktion f und einer ebenen Welle ebenfalls holomorph ist, kann das Lemma 4.24 angewendet werden. Es ergibt sich

$$\begin{aligned}\|v-\mathfrak{I}_{I_\ell}[v]\|_{\infty,{}^{I_\ell}D_\varrho} &\leq c_{q,\varrho}q^m\|v\|_{\infty,{}^{I_{\ell-1}}D_\varrho} = c_{q,\varrho}q^m\|e^{-\mathrm{i}\kappa\dot{c}_\ell^i\cdot}f\|_{\infty,{}^{I_{\ell-1}}D_\varrho}\\ &= c_{q,\varrho}q^m\|e^{-\mathrm{i}\kappa\dot{c}_\ell^i\cdot}e^{\mathrm{i}\kappa\dot{c}_0^i\cdot}e^{-\mathrm{i}\kappa\dot{c}_0^i\cdot}f\|_{\infty,{}^{I_{\ell-1}}D_\varrho}\\ &\leq c_{q,\varrho}q^m\|e^{\mathrm{i}\kappa(\dot{c}_0^i-\dot{c}_\ell^i)\cdot}\|_{\infty,{}^{I_{\ell-1}}D_\varrho}\|e^{-\mathrm{i}\kappa\dot{c}_0^i\cdot}f\|_{\infty,{}^{I_{\ell-1}}D_\varrho}\\ &\overset{(4.3.11)}{\leq} e^{\gamma_\varrho}c_{q,\varrho}q^m\|e^{-\mathrm{i}\kappa\dot{c}_0^i\cdot}f\|_{\infty,{}^{I_{\ell-1}}D_\varrho}.\end{aligned}$$

Für den letzten Teil der Behauptung füge zunächst eine nahrhafte Null ein und ziehe die Norm auseinander

$$\begin{aligned}\|e^{-\mathrm{i}\kappa\dot{c}_0^i\cdot}\mathfrak{I}_{I_\ell}^{i,c_\ell}[f]\|_{\infty,{}^{I_\ell}D_\varrho} &= \|e^{-\mathrm{i}\kappa\dot{c}_0^i\cdot}f-e^{-\mathrm{i}\kappa\dot{c}_0^i\cdot}f+e^{-\mathrm{i}\kappa\dot{c}_0^i\cdot}\mathfrak{I}_{I_\ell}^{i,c_\ell}[f]\|_{\infty,{}^{I_\ell}D_\varrho}\\ &\overset{\Delta}{\leq} \|e^{-\mathrm{i}\kappa\dot{c}_0^i\cdot}f\|_{\infty,{}^{I_\ell}D_\varrho}+\|e^{-\mathrm{i}\kappa\dot{c}_0^i\cdot}f-e^{-\mathrm{i}\kappa\dot{c}_0^i\cdot}\mathfrak{I}_{I_\ell}^{i,c_\ell}[f]\|_{\infty,{}^{I_\ell}D_\varrho}.\end{aligned}$$

Für die zweite Norm nutze (4.3.11)

$$\begin{aligned}&\|e^{-\mathrm{i}\kappa\dot{c}_0^i\cdot}f-e^{-\mathrm{i}\kappa\dot{c}_0^i\cdot}\mathfrak{I}_{I_\ell}^{i,c_\ell}[f]\|_{\infty,{}^{I_\ell}D_\varrho} \leq \|e^{\mathrm{i}\kappa(\dot{c}_\ell^i-\dot{c}_0^i)\cdot}\left(e^{-\mathrm{i}\kappa\dot{c}_\ell^i\cdot}f-\mathfrak{I}_{I_\ell}[e^{-\mathrm{i}\kappa\dot{c}_\ell^i\cdot}f]\right)\|_{\infty,{}^{I_\ell}D_\varrho}\\ &\leq \|e^{\mathrm{i}\kappa(\dot{c}_\ell^i-\dot{c}_0^i)\cdot}\|_{\infty,{}^{I_\ell}D_\varrho}\|e^{-\mathrm{i}\kappa\dot{c}_\ell^i\cdot}f-\mathfrak{I}_{I_\ell}[e^{-\mathrm{i}\kappa\dot{c}_\ell^i\cdot}f]\|_{\infty,{}^{I_\ell}D_\varrho}\\ &\leq e^{\gamma_\varrho}\|e^{-\mathrm{i}\kappa\dot{c}_\ell^i\cdot}f-\mathfrak{I}_{I_\ell}[e^{-\mathrm{i}\kappa\dot{c}_\ell^i\cdot}f]\|_{\infty,{}^{I_\ell}D_\varrho},\end{aligned}$$

um dann auf Lemma 4.24 zurück greifen zu können

$$\begin{aligned}e^{\gamma_\varrho}\|e^{-\mathrm{i}\kappa\dot{c}_\ell^i\cdot}f-\mathfrak{I}_{I_\ell}[e^{-\mathrm{i}\kappa\dot{c}_\ell^i\cdot}f]\|_{\infty,{}^{I_\ell}D_\varrho} &\leq e^{\gamma_\varrho}c_{q,\varrho}q^m\|e^{-\mathrm{i}\kappa\dot{c}_\ell^i\cdot}f\|_{\infty,{}^{I_{\ell-1}}D_\varrho}\\ &= e^{\gamma_\varrho}c_{q,\varrho}q^m\|e^{\mathrm{i}\kappa(\dot{c}_0^i-\dot{c}_\ell^i)\cdot}e^{-\mathrm{i}\kappa\dot{c}_0^i\cdot}f\|_{\infty,{}^{I_{\ell-1}}D_\varrho}\\ &\leq e^{2\gamma_\varrho}c_{q,\varrho}q^m\|e^{-\mathrm{i}\kappa\dot{c}_0^i\cdot}f\|_{\infty,{}^{I_{\ell-1}}D_\varrho}.\end{aligned}$$

Durch das Zusammenfügen der Teilabschätzungen folgt die Behauptung

$$\begin{aligned}\|e^{-\mathrm{i}\kappa\dot{c}_0^i\cdot}\mathfrak{I}_{I_\ell}^{i,c_\ell}[f]\|_{\infty,{}^{I_\ell}D_\varrho} &\leq \|e^{-\mathrm{i}\kappa\dot{c}_0^i\cdot}f\|_{\infty,{}^{I_\ell}D_\varrho} + e^{2\gamma_\varrho}c_{q,\varrho}q^m\|e^{-\mathrm{i}\kappa\dot{c}_0^i\cdot}f\|_{\infty,{}^{I_{\ell-1}}D_\varrho}\\ &\leq \|e^{-\mathrm{i}\kappa\dot{c}_0^i\cdot}f\|_{\infty,{}^{I_{\ell-1}}D_\varrho} + e^{2\gamma_\varrho}c_{q,\varrho}q^m\|e^{-\mathrm{i}\kappa\dot{c}_0^i\cdot}f\|_{\infty,{}^{I_{\ell-1}}D_\varrho}\\ &= (1+e^{2\gamma_\varrho}c_{q,\varrho}q^m)\|e^{-\mathrm{i}\kappa\dot{c}_0^i\cdot}f\|_{\infty,{}^{I_{\ell-1}}D_\varrho}.\end{aligned}$$

□

Bemerkung 25 : *Die Aussagen aus Lemma 4.26 bleiben auch erhalten, wenn statt $\dot{c}_0^i$ eine andere Richtung $\dot{c}_k^i$ für $k \in \underline{\ell_\rfloor}$ aus der Sequenz an Richtungen betrachtet wird.*

Mit diesem Wissen kann eine Aussage zum Fehler der Reinterpolation bewiesen werden. Der Beweis orientiert sich am Grundgerüst des Beweises [3, Thm 5.6], nutzt aber das vorherige Lemma für die Abschätzungen der einzelnen Schritte.

Satz 4.27 (Fehler der Reinterpolation)
Seien für $L \in \underline{p_{\mathcal{I}\rfloor}}$ Sequenzen von Clustern $\mathfrak{s}_t^L$, $\mathfrak{s}_s^L$, deren dazugehörige Gebiete die Kontraktionseigenschaft (4.3.5) *erfüllen, und eine passende Sequenz an Richtungen $\mathfrak{s}_c^L$ gegeben. Weiter seien ein $\varrho \in \mathbb{R}_{>1}$ und $\sigma := \vartheta(\varrho)$ gegeben, der betrachtete Interpolationsoperator von Ordnung $m \in \mathbb{N}_0$ sei stabil nach Bemerkung 6 und es existiere ein $\gamma_\varrho \in \mathbb{R}_{>0}$, das* (4.3.10) *erfüllt. Dann gibt es zu jedem $q \in (\sigma^{-1}, 1)$ ein $c_{q,\varrho}$ wie in Lemma 4.24, mit dem für alle $i \in \underline{6_\rfloor}$ und für alle auf ${}^{I_{0,i}}D_\varrho$ holomorphen Funktionen f*

$$\left\|e^{\mathrm{i}\kappa\dot{c}_0^i\cdot}f - (\mathfrak{I}_{I_L}^{i,c_L}\circ\cdots\circ\mathfrak{I}_{I_1}^{i,c_1})[e^{\mathrm{i}\kappa\dot{c}_0^i\cdot}f]\right\|_{\infty,I_L} \leq e^{-\gamma_\varrho}\left((1+e^{2\gamma_\varrho}c_{q,\varrho}q^m)^L - 1\right)\|f\|_{\infty,{}^{I_{0,i}}D_\varrho}$$

gilt. Außerdem existieren zu jedem festen $\widetilde{q} \in (q,1)$ ein $\mathcal{C}_{\varrho,q} > 0$, welches von $\widetilde{q}$, der Wahl von Zwischengrößen in $(q,\widetilde{q})$, γ_ϱ und $c_{q,\varrho}$ abhängt, so dass für

$$m \geq \mathcal{C}_{\varrho,q}(1+\ln(L)) \qquad \Rightarrow \qquad e^{-\gamma_\varrho}\left((1+e^{2\gamma_\varrho}c_{q,\varrho}q^m)^L - 1\right) \leq \widetilde{q}^m$$

erfüllt ist.

Beweis: Seien ein $i \in \underline{6_\rfloor}$ und eine auf ${}^{I_{0,i}}D_\varrho$ holomorphe Funktion f gegeben. Nutze eine Teleskopsumme, um die Fehlerabschätzung aus der einmaligen Interpolation zu übertragen. Mit der Teleskopsumme

$$\begin{aligned}T_1 :&= Id - \mathfrak{I}_{I_L}^{i,c_L}\circ\cdots\circ\mathfrak{I}_{I_1}^{i,c_1}\\ &= (Id-\mathfrak{I}_{I_1}^{i,c_1}) + (Id-\mathfrak{I}_{I_2}^{i,c_2})\circ\mathfrak{I}_{I_1}^{i,c_1} + \cdots + (Id-\mathfrak{I}_{I_L}^{i,c_L})\circ\mathfrak{I}_{I_{L-1}}^{i,c_{L-1}}\circ\cdots\circ\mathfrak{I}_{I_1}^{i,c_1}\end{aligned} \tag{4.3.12}$$

zusammen mit Lemma 4.26 ergibt sich

$$\begin{aligned}\left\|T_1[e^{\mathrm{i}\kappa\dot{c}_0^i\cdot}f]\right\|_{\infty,I_L} &= \left\|\sum_{\ell=1}^{L}(Id-\mathfrak{I}_{I_\ell}^{i,c_\ell})(\mathfrak{I}_{I_{\ell-1}}^{i,c_{\ell-1}}\circ\cdots\circ\mathfrak{I}_{I_1}^{i,c_1})[e^{\mathrm{i}\kappa\dot{c}_0^i\cdot}f]\right\|_{\infty,I_L}\\ &\overset{\Delta}{\leq}\sum_{\ell=1}^{L}\left\|(Id-\mathfrak{I}_{I_\ell}^{i,c_\ell})(\mathfrak{I}_{I_{\ell-1}}^{i,c_{\ell-1}}\circ\cdots\circ\mathfrak{I}_{I_1}^{i,c_1})[e^{\mathrm{i}\kappa\dot{c}_0^i\cdot}f]\right\|_{\infty,I_\ell}\\ &\leq e^{\gamma_\varrho}c_{q,\varrho}q^m\sum_{\ell=1}^{L}\left\|e^{-\mathrm{i}\kappa\dot{c}_0^i\cdot}(\mathfrak{I}_{I_{\ell-1}}^{i,c_{\ell-1}}\circ\cdots\circ\mathfrak{I}_{I_1}^{i,c_1})[e^{\mathrm{i}\kappa\dot{c}_0^i\cdot}f]\right\|_{\infty,{}^{I_{\ell-1}}D_\varrho}.\end{aligned}$$

Bei der verbliebenen Summe kann die letzte Abschätzung aus dem Lemma 4.26 angewendet werden. Dies führt zu

$$\begin{aligned}\left\|T_1[e^{\mathrm{i}\kappa\dot{c}_0^i\cdot}f]\right\|_{\infty,I_L} &\leq e^{\gamma_\varrho}c_{q,\varrho}q^m\sum_{\ell=1}^{L}(1+e^{2\gamma_\varrho}c_{q,\varrho}q^m)^{\ell-1}\left\|e^{-\mathrm{i}\kappa\dot{c}_0^i\cdot}e^{\mathrm{i}\kappa\dot{c}_0^i\cdot}f\right\|_{\infty,{}^{I_{0,i}}D_\varrho}\\ &= e^{\gamma_\varrho}c_{q,\varrho}q^m\sum_{\ell=1}^{L}(1+e^{2\gamma_\varrho}c_{q,\varrho}q^m)^{\ell-1}\left\|f\right\|_{\infty,{}^{I_{0,i}}D_\varrho}.\end{aligned}$$

Mit einer Indexverschiebung und der Partialsumme der geometrischen Reihe ergibt sich

$$\begin{aligned}e^{\gamma_\varrho}c_{q,\varrho}q^m\sum_{\ell=1}^{L}(1+e^{2\gamma_\varrho}c_{q,\varrho}q^m)^{\ell-1} &= e^{\gamma_\varrho}c_{q,\varrho}q^m\sum_{\ell=0}^{L-1}(1+e^{2\gamma_\varrho}c_{q,\varrho}q^m)^{\ell}\\ &= e^{\gamma_\varrho}c_{q,\varrho}q^m\tfrac{(1+e^{2\gamma_\varrho}c_{q,\varrho}q^m)^L-1}{e^{2\gamma_\varrho}c_{q,\varrho}q^m}\\ &= e^{-\gamma_\varrho}\left((1+e^{2\gamma_\varrho}c_{q,\varrho}q^m)^L-1\right),\end{aligned}$$

was den ersten Teil der Behauptung zeigt.
Für den zweiten Teil seien ein $\widetilde{q}\in(q,1)$ und ein $\widehat{q}\in(q,\widetilde{q})$ gegeben. Die exponentielle Abnahme von q^m zusammen mit der Tatsache, dass $e^{2\gamma_\varrho}$ unabhängig von m und L beschränkt ist, liefert die Existenz eines $\widehat{m}\in\mathbb{N}$, so dass

$$e^{2\gamma_\varrho}c_{q,\varrho}q^m<1 \qquad \text{für alle } m\geq\widehat{m}$$

gilt. Mit der gleichen Argumentation existiert dann auch zu $\widehat{q}$ ein $m_{\varrho,q}\geq\widehat{m}$, so dass

$$e^{2\gamma_\varrho}c_{q,\varrho}q^m\widehat{q}^{-m}<1 \qquad \text{für alle } m\geq m_{\varrho,q}$$

erfüllt ist. Dies wiederum führt für $m\geq m_{\varrho,q}$ zu

$$(1+e^{2\gamma_\varrho}c_{q,\varrho}q^m)^L-1\leq(1+\widehat{q}^m)^L-1=(1+\widehat{q}^m)^{\widehat{q}^{-m}\widehat{q}^mL}-1.$$

Das monotone Wachstum der Folge $\left\{\left(1+\frac{1}{n}\right)^n\right\}_{n\in\mathbb{N}}$ mit ihrem Grenzwert e ermöglicht die Abschätzung $\sup\limits_{x\in(0,1]}(1+x)^{\frac{1}{x}}\leq e$. Dies führt zu

$$(1+e^{2\gamma_\varrho}c_{q,\varrho}q^m)^L-1\leq e^{\widehat{q}^mL}-1.$$

Grundsätzlich gilt mit $\widehat{q} < 1$ und $\ln(L) \geq 0$

$$\widehat{q}^m L \leq 1 \qquad \Longleftrightarrow \qquad m \ln(\widehat{q}) \leq 0 - \ln(L) \qquad \Longleftrightarrow$$
$$m \geq \tfrac{1}{|\ln(\widehat{q})|}(\ln(L)).$$

Damit auch $m \geq 1$ erfüllt ist, reicht es mit der eben gemachten Betrachtung aus, $m \geq \frac{1}{|\ln(\widehat{q})|}(1 + \ln(L))$ zu fordern. Dazu wähle $\mathcal{C}_{\varrho,q} \geq \frac{1}{|\ln(\widehat{q})|}$ und $m \geq \mathcal{C}_{\varrho,q}(1 + \ln(L))$. Auf diese Weise ist durch die Wahl von m und $\mathcal{C}_{\varrho,q}$ sichergestellt, dass $L\widehat{q}^m \leq 1$ erfüllt ist. Mit dem zusätzlichen Wissen, dass $e^x - 1 \leq ex$ für alle $x \in [0, 1]$ gilt[vi], ergibt sich

$$\begin{aligned}(\widetilde{q})^{-m}\left((1 + e^{2\gamma_\varrho} c_{q,\varrho} q^m)^L - 1\right) \leq (\widetilde{q})^{-m}(e^{\widehat{q}^m L} - 1) &\leq e\left(\frac{\widetilde{q}}{\widehat{q}}\right)^{-m} L \\ &= e^{\ln(e)+\ln(L)-m\ln\left(\frac{\widetilde{q}}{\widehat{q}}\right)} \\ &\leq e^{\ln(e)+\ln(L)-\mathcal{C}_{\varrho,q}(1+\ln(L))\ln\left(\frac{\widetilde{q}}{\widehat{q}}\right)}.\end{aligned}$$

Mit einer endgültigen Wahl von

$$\mathcal{C}_{\varrho,q} := \max\left\{ m_{\varrho,q}, \tfrac{1}{|\ln(\widehat{q})|}, \frac{1}{\ln\left(\frac{\widetilde{q}}{\widehat{q}}\right)} \right\}$$

ergibt sich dann

$$\begin{aligned}(\widetilde{q})^{-m}\left((1 + e^{2\gamma_\varrho} c_{q,\varrho} q^m)^L - 1\right) &\leq e^{\ln(e)+\ln(L)-(1+\ln(L))} \\ &= e^{\ln(e)-1} = e^0 = 1.\end{aligned}$$

Entsprechend konvergiert der Term $(1 + e^{2\gamma_\varrho} c_{q,\varrho} q^m)^L - 1$ für $m \to \infty$ gegen null.
Der verbliebene Faktor $e^{-\gamma_\varrho}$ hängt weder von der Länge L der Clustersequenz noch von m ab. Zudem gilt $\gamma_\varrho > 0$, so dass $e^{\gamma_\varrho} \geq 1$ erfüllt ist, entsprechend folgt $e^{-\gamma_\varrho} \leq 1$, so dass direkt die Behauptung folgt

$$e^{-\gamma_\varrho}\left((1 + e^{2\gamma_\varrho} c_{q,\varrho} q^m)^L - 1\right) \leq \left((1 + e^{2\gamma_\varrho} c_{q,\varrho} q^m)^L - 1\right) \leq \widetilde{q}^m.$$

□

Um eine Stabilitätskonstante für den Interpolationsoperator $\mathfrak{I}_I^{i,\mathfrak{S}_c^L}$ herzuleiten, wird auf ähnliche Weise wie in [9] vorgegangen. Dazu ist zunächst noch eine Abschätzung für die Exponentialfunktion, welche in in der folgenden Bemerkung zu finden ist, notwendig.

Bemerkung 26 (Abschätzung Exponentialfunktion): *Für alle $x \in \mathbb{R}$ gilt*

$$1 + x \leq e^x.$$

[vi] Leicht durch Taylor-Entwicklung in 0 um x zu zeigen.

Beweis: Die Aussage lässt sich durch eine Taylor-Entwicklung der Exponentialfunktion um null zeigen. Es gilt für ein η zwischen null und x

$$e^x = e^0 + e^0 x + \frac{e^\eta}{2}x^2 = 1 + x + \frac{e^\eta}{2}x^2.$$

Da unabhängig vom Vorzeichen von x und η gilt, dass $e^\eta > 0$ und $x^2 > 0$, folgt

$$e^x = 1 + x + \frac{e^\eta}{2}x^2 \geq 1 + x$$

und damit die Behauptung. □

Damit ist alles beisammen, um die Stabilität der geschachtelten richtungsabhängigen Interpolation nachzuweisen.

Lemma 4.28 (Stabilität der Reinterpolation)
Seien die Voraussetzungen von Satz 4.27 erfüllt und ein $p \in (q, 1]$ gegeben. Dann gilt für alle $i \in \underline{6}$ und

$$m \geq \frac{\ln(L)}{\ln\left(\frac{p}{q}\right)},$$

dass der geschachtelte richtungsabhängige Interpolationsoperator

$$\left\|\mathfrak{I}_I^{i,\mathfrak{s}_c^L}\right\|_{op,C(I_L)\leftarrow C(I_0)} \leq \Lambda_m(p^m c_{q,\varrho}\widehat{c}_{\gamma_\varrho,q,\varrho} + 1)$$

erfüllt, wobei $\widehat{c}_{\gamma_\varrho,q,\varrho} = e^{\gamma_\varrho}e^{e^{2\gamma_\varrho}c_{q,\varrho}}$ eine Konstante ist.

Beweis: Seien ein $i \in \underline{6}$ und eine Funktion $v_i \in C(I_{0,i})$ gegeben. Definiere ein Polynom $p_{1,i} \in \Pi_m$ durch $p_{1,i} := \mathfrak{I}_{I_{1,i}}[e^{-\mathrm{i}\kappa\dot{c}_1^i\cdot}v_i]$. Vorweg halte fest, dass dann

$$\mathfrak{I}_{I_1}^{i,c_1}[v_i] = e^{\mathrm{i}\kappa\dot{c}_1^i\cdot}p_{1,i}$$

gilt. Da der Interpolationsoperator auf den Raum der Polynome abbildet, kann die Norm von $p_{1,i}$ mit Hilfe der Bernstein-Ungleichung (4.1.10) abgeschätzt werden

$$\begin{aligned}
\|p_{1,i}\|_{\infty,{}^{I_1}D_\varrho} &= \|\mathfrak{I}_{[-1,1]}[(e^{-\mathrm{i}\kappa\dot{c}_1^i\cdot}v_i)\circ\Phi_{I_1}]\|_{\infty,D_\varrho} \\
&\leq \varrho^m\|\mathfrak{I}_{[-1,1]}[(e^{-\mathrm{i}\kappa\dot{c}_1^i\cdot}v_i)\circ\Phi_{I_1}]\|_{\infty,[-1,1]} \\
&= \varrho^m\|\mathfrak{I}_{I_1}[e^{-\mathrm{i}\kappa\dot{c}_1^i\cdot}v_i]\|_{\infty,I_1} \\
&\leq \varrho^m\Lambda_m\|e^{-\mathrm{i}\kappa\dot{c}_1^i\cdot}v_i\|_{\infty,I_0} \overset{(4.2.1)}{=} \varrho^m\Lambda_m\|v_i\|_{\infty,I_0}.
\end{aligned}$$

Weiter definiere Funktionen h_ℓ für $\ell \in \underline{L-1}$ mit

$$h_\ell(x) := \begin{cases} e^{-\mathrm{i}\kappa\dot{c}_2^i x}e^{\mathrm{i}\kappa\dot{c}_1^i x}p_{1,i}(x) & \text{für } \ell = 1, \\ e^{-\mathrm{i}\kappa\dot{c}_{\ell+1}^i x}\mathfrak{I}_{I_\ell}^{i,c_\ell}\circ\cdots\circ\mathfrak{I}_{I_2}^{i,c_2}[e^{\mathrm{i}\kappa\dot{c}_1^i x}p_{1,i}](x) & \text{sonst.} \end{cases}$$

Mit h_{L-1} kann zunächst

$$\begin{aligned}\left\|\mathfrak{I}_I^{i,\mathfrak{s}_c^L}[v_i]\right\|_{\infty,I_L} &= \left\|\mathfrak{I}_{I_L}^{i,c_L}\circ\cdots\circ\mathfrak{I}_{I_2}^{i,c_2}[e^{\mathrm{i}\kappa\dot{c}_1^i\cdot}p_{1,i}]\right\|_{\infty,I_L} = \left\|e^{\mathrm{i}\kappa\dot{c}_L^i\cdot}\mathfrak{I}_{I_L}[h_{L-1}]\right\|_{\infty,I_L}\\ &= \|\mathfrak{I}_{I_L}[h_{L-1}]\|_{\infty,I_L} \overset{\Delta}{\leq} \|h_{L-1}-\mathfrak{I}_{I_L}[h_{L-1}]\|_{\infty,I_L} + \|h_{L-1}\|_{\infty,I_L}\end{aligned}$$

erhalten werden, um dann beim zweiten Summanden mit h_{L-2} weiter umzuformen

$$\begin{aligned}\|h_{L-1}\|_{\infty,I_L} &\overset{(4.2.1)}{=} \left\|\mathfrak{I}_{I_{L-1}}^{i,c_{L-1}}\circ\cdots\circ\mathfrak{I}_{I_2}^{i,c_2}[e^{\mathrm{i}\kappa\dot{c}_1^i\cdot}p_{1,i}]\right\|_{\infty,I_L}\\ &\leq \left\|\mathfrak{I}_{I_{L-1}}^{i,c_{L-1}}\circ\cdots\circ\mathfrak{I}_{I_2}^{i,c_2}[e^{\mathrm{i}\kappa\dot{c}_1^i\cdot}p_{1,i}]\right\|_{\infty,I_{L-1}}\\ &= \left\|e^{\mathrm{i}\kappa\dot{c}_{L-1}^i\cdot}\mathfrak{I}_{I_{L-1}}[h_{L-2}]\right\|_{\infty,I_{L-1}}\\ &\overset{\Delta}{\leq} \left\|h_{L-2}-\mathfrak{I}_{I_{L-1}}[h_{L-2}]\right\|_{\infty,I_{L-1}} + \|h_{L-2}\|_{\infty,I_{L-1}}.\end{aligned}$$

Ebenso kann mit der Norm von h_{L-2} verfahren werden, so dass sich insgesamt

$$\begin{aligned}\left\|\mathfrak{I}_I^{i,\mathfrak{s}_c^L}[v_i]\right\|_{\infty,I_L} &\leq \|h_{L-1}-\mathfrak{I}_{I_L}[h_{L-1}]\|_{\infty,I_L} + \|h_{L-1}\|_{\infty,I_L}\\ &\leq \sum_{\ell=2}^{L}\|h_{\ell-1}-\mathfrak{I}_{I_\ell}[h_{\ell-1}]\|_{\infty,I_\ell} + \|h_1\|_{\infty,I_2}\end{aligned}$$

ergibt. Die Norm von h_1 kann durch

$$\begin{aligned}\|h_1\|_{\infty,I_2} &= \left\|e^{-\mathrm{i}\kappa\dot{c}_2^i\cdot}e^{\mathrm{i}\kappa\dot{c}_1^i\cdot}p_{1,i}\right\|_{\infty,I_2} \overset{(4.2.1)}{=} \|p_{1,i}\|_{\infty,I_2} \leq \|p_{1,i}\|_{\infty,I_1}\\ &\leq \Lambda_m\|v_i\|_{\infty,I_0}\end{aligned}$$

abgeschätzt werden, entsprechend folgt

$$\left\|\mathfrak{I}_I^{i,\mathfrak{s}_c^L}[v_i]\right\|_{\infty,I_L} \leq \sum_{\ell=2}^{L}\|h_{\ell-1}-\mathfrak{I}_{I_\ell}[h_{\ell-1}]\|_{\infty,I_\ell} + \Lambda_m\|v_i\|_{\infty,I_0}.$$

Mit den verbliebenen Termen in der Summe kann wie folgt verfahren werden.
Sei ein $\ell \in \underline{L_\sqcup}_2$ gegeben und nutze die Bemerkung 23 jedoch mit ϱ in der Rolle von τ und einer Wahl von $\varrho = 1$, dann ergibt sich

$$\begin{aligned}\|h_{\ell-1}-\mathfrak{I}_{I_\ell}[h_{\ell-1}]\|_{\infty,I_\ell} &= \|h_{\ell-1}-\mathfrak{I}_{I_\ell}[h_{\ell-1}]\|_{\infty,{}^{I_\ell}D_1} \leq c_{q,\varrho}q^m\varrho^{-m}\|h_{\ell-1}\|_{\infty,{}^{I_\ell}D_{\sigma\varrho 1}}\\ &= c_{q,\varrho}q^m\varrho^{-m}\|h_{\ell-1}\|_{\infty,{}^{I_\ell}D_{\sigma\varrho}} \leq c_{q,\varrho}q^m\varrho^{-m}\|h_{\ell-1}\|_{\infty,{}^{I_{\ell-1}}D_\varrho}.\end{aligned}$$

Schreibe $h_{\ell-1}$ nun wieder aus und nutze die Stabilität des richtungsabhängigen Interpolationsoperators. Nach Bemerkung 25 kann die Stabilitätsaussage aus Lemma 4.26 auch dann

angewendet werden, wenn eine ebene Welle mit Richtung $\dot{c}_1^i$ genutzt wird, so dass sich durch mehrfaches Ausnutzen der Stabilität

$$\begin{aligned}
\|h_{\ell-1}\|_{\infty,{}^{I_{\ell-1}}D_\varrho} &= \left\| e^{-\mathrm{i}\kappa\dot{c}_\ell^i\cdot}\mathfrak{I}_{I_{\ell-1}}^{i,c_{\ell-1}} \circ \cdots \circ \mathfrak{I}_{I_2}^{i,c_2}[e^{\mathrm{i}\kappa\dot{c}_1^i\cdot}p_{1,i}] \right\|_{\infty,{}^{I_{\ell-1}}D_\varrho} \\
&\leq e^{\gamma_\varrho} \left\| e^{-\mathrm{i}\kappa\dot{c}_1^i\cdot}\mathfrak{I}_{I_{\ell-1}}^{i,c_{\ell-1}} \circ \cdots \circ \mathfrak{I}_{I_2}^{i,c_2}[e^{\mathrm{i}\kappa\dot{c}_1^i\cdot}p_{1,i}] \right\|_{\infty,{}^{I_{\ell-1}}D_\varrho} \\
&\overset{4.26}{\leq} e^{\gamma_\varrho}(1+e^{2\gamma_\varrho}c_{q,\varrho}q^m) \left\| e^{-\mathrm{i}\kappa\dot{c}_1^i\cdot}\mathfrak{I}_{I_{\ell-2}}^{i,c_{\ell-2}} \circ \cdots \circ \mathfrak{I}_{I_2}^{i,c_2}[e^{\mathrm{i}\kappa\dot{c}_1^i\cdot}p_{1,i}] \right\|_{\infty,{}^{I_{\ell-2}}D_\varrho} \\
&\leq e^{\gamma_\varrho}(1+e^{2\gamma_\varrho}c_{q,\varrho}q^m)^{\ell-2} \left\| e^{-\mathrm{i}\kappa\dot{c}_1^i\cdot}e^{\mathrm{i}\kappa\dot{c}_1^i\cdot}p_{1,i} \right\|_{\infty,{}^{I_1}D_\varrho} \\
&= e^{\gamma_\varrho}(1+e^{2\gamma_\varrho}c_{q,\varrho}q^m)^{\ell-2} \|p_{1,i}\|_{\infty,{}^{I_1}D_\varrho}
\end{aligned}$$

ergibt. Insgesamt folgt für den betrachteten Summanden

$$\begin{aligned}
\|h_{\ell-1} - \mathfrak{I}_{I_\ell}[h_{\ell-1}]\|_{\infty,I_\ell} &\leq c_{q,\varrho}q^m\varrho^{-m}e^{\gamma_\varrho}(1+e^{2\gamma_\varrho}c_{q,\varrho}q^m)^{\ell-2}\|p_{1,i}\|_{\infty,{}^{I_1}D_\varrho} \\
&\leq c_{q,\varrho}q^m\varrho^{-m}e^{\gamma_\varrho}(1+e^{2\gamma_\varrho}c_{q,\varrho}q^m)^{\ell-2}\varrho^m\Lambda_m\|v_i\|_{\infty,I_0} \\
&= c_{q,\varrho}q^m e^{\gamma_\varrho}(1+e^{2\gamma_\varrho}c_{q,\varrho}q^m)^{\ell-2}\Lambda_m\|v_i\|_{\infty,I_0}.
\end{aligned}$$

Bevor die Summe weiter aufgelöst wird, stelle fest, dass für $p \in (q,1]$ die Bedingung

$$m \geq \frac{\ln(L)}{\ln\left(\frac{p}{q}\right)}$$

äquivalent zu

$$\begin{aligned}
&m\ln\left(\frac{p}{q}\right) \geq \ln(L) \iff 0 \geq \ln(L) - m\ln\left(\frac{p}{q}\right) \iff 0 \geq \ln(L) + m\ln\left(\frac{q}{p}\right) \\
&\iff e^0 \geq e^{\ln(L)}e^{m\ln\left(\frac{q}{p}\right)} \iff 1 \geq L\left(\frac{q}{p}\right)^m
\end{aligned}$$

ist. Da $qp^{-1} \geq q$ gilt, folgt auch

$$1 \geq Lq^m.$$

Entsprechend gilt auch für alle $\ell \in \underline{L}_{\rfloor 2}$ dann $1 \geq \ell q^m$. Zusammen mit Bemerkung 26 ergibt sich für

$$(1+e^{2\gamma_\varrho}c_{q,\varrho}q^m)^{\ell-2} \leq e^{e^{2\gamma_\varrho}c_{q,\varrho}q^m(\ell-2)} \leq e^{e^{2\gamma_\varrho}c_{q,\varrho}},$$

so dass

$$e^{\gamma_\varrho}(1+e^{2\gamma_\varrho}c_{q,\varrho}q^m)^{\ell-2} \leq e^{\gamma_\varrho}e^{e^{2\gamma_\varrho}c_{q,\varrho}} = \widehat{c}_{\gamma_\varrho,q,\varrho}$$

folgt. Somit kann jeder Summand auch mit

$$\begin{aligned}\|h_{\ell-1} - \mathfrak{I}_{I_\ell}[h_{\ell-1}]\|_{\infty,I_\ell} &\leq c_{q,\varrho} q^m e^{\gamma_\varrho} (1 + e^{2\gamma_\varrho} c_{q,\varrho} q^m)^{\ell-2} \Lambda_m \|v_i\|_{\infty,I_0} \\ &\leq c_{q,\varrho} q^m \widehat{c}_{\gamma_\varrho,q,\varrho} \Lambda_m \|v_i\|_{\infty,I_0}\end{aligned}$$

beschränkt werden. Es folgt

$$\begin{aligned}\sum_{\ell=2}^{L} \|h_{\ell-1} - \mathfrak{I}_{I_\ell}[h_{\ell-1}]\|_{\infty,I_\ell} &\leq \sum_{\ell=2}^{L} c_{q,\varrho} q^m \widehat{c}_{\gamma_\varrho,q,\varrho} \Lambda_m \|v_i\|_{\infty,I_0} \\ &= (L-1) c_{q,\varrho} q^m \widehat{c}_{\gamma_\varrho,q,\varrho} \Lambda_m \|v_i\|_{\infty,I_0}\end{aligned}$$

und damit insgesamt

$$\left\| \mathfrak{I}_I^{i,\mathfrak{s}_c^L}[v_i] \right\|_{\infty,I_L} \leq (L-1) c_{q,\varrho} q^m \widehat{c}_{\gamma_\varrho,q,\varrho} \Lambda_m \|v_i\|_{\infty,I_0} + \Lambda_m \|v_i\|_{\infty,I_0}.$$

Für $p \in (q, 1]$ und mit der Bedingung

$$m \geq \frac{\ln(L)}{\ln\left(\frac{p}{q}\right)}$$

ergibt sich

$$\begin{aligned}(L-1) c_{q,\varrho} q^m \widehat{c}_{\gamma_\varrho,q,\varrho} \Lambda_m \|v_i\|_{\infty,I_0} &\leq (L-1)(qp^{-1})^m p^m c_{q,\varrho} \widehat{c}_{\gamma_\varrho,q,\varrho} \Lambda_m \|v_i\|_{\infty,I_0} \\ &\leq p^m c_{q,\varrho} \widehat{c}_{\gamma_\varrho,q,\varrho} \Lambda_m \|v_i\|_{\infty,I_0}.\end{aligned}$$

Insgesamt folgt so

$$\begin{aligned}\left\| \mathfrak{I}_I^{i,\mathfrak{s}_c^L}[v_i] \right\|_{\infty,I_L} &\leq p^m c_{q,\varrho} \widehat{c}_{\gamma_\varrho,q,\varrho} \Lambda_m \|v_i\|_{\infty,I_0} + \Lambda_m \|v_i\|_{\infty,I_0} \\ &\leq \Lambda_m (p^m c_{q,\varrho} \widehat{c}_{\gamma_\varrho,q,\varrho} + 1) \|v_i\|_{\infty,I_0}\end{aligned}$$

und damit für die Operatornorm

$$\left\| \mathfrak{I}_I^{i,\mathfrak{s}_c^L} \right\|_{op,C(I_L)\leftarrow C(I_0)} \leq \Lambda_m (p^m c_{q,\varrho} \widehat{c}_{\gamma_\varrho,q,\varrho} + 1).$$

□

Schließlich kann mit Hilfe der Stabilität der Reinterpolation im Eindimensionalen auch eine Aussagen für den mehrdimensionalen Fall getroffen werden.

Theorem 4.29 (Stabilität mehrdimensionale Reinterpolation)
Seien für $L \in \underline{p_{\mathcal{I}}}$ Sequenzen an Clustern $\mathfrak{s}_t^L$, $\mathfrak{s}_s^L$, deren dazugehörige Gebiete die Kontraktionseigenschaft (4.3.5) *erfüllen, und eine passende Sequenz an Richtungen $\mathfrak{s}_c^L$ gegeben. Weiter seien ein $\varrho \in \mathbb{R}_{>1}$ und $\sigma := \vartheta(\varrho)$ gegeben, der betrachtete Interpolationsoperator*

sei stabil nach Bemerkung 6 und es existiere ein $\gamma_\varrho \in \mathbb{R}_{>0}$, das (4.3.10) *erfüllt. Dann gilt für jedes $q \in (\sigma^{-1}, 1)$, $p \in (q, 1]$ und jede Interpolationsordnung*

$$m \geq \frac{\ln(L)}{\ln\left(\frac{p}{q}\right)},$$

für die Operatornorm der Reinterpolation im Mehrdimensionalen

$$\left\| \mathfrak{I}^{c_L}_{Q_{t_L} \times Q_{s_L}} \circ \cdots \circ \mathfrak{I}^{c_1}_{Q_{t_1} \times Q_{s_1}} \right\|_{op, C(Q_{t_L} \times Q_{s_L}) \leftarrow C(Q_{t_0} \times Q_{s_0})} \leq \Lambda_m^6 (p^m c_{q,\varrho} \widehat{c}_{\gamma_\varrho, q, \varrho} + 1)^6,$$

wobei $\widehat{c}_{\gamma_\varrho, q, \varrho}$ die Konstante aus Lemma 4.28 ist.

Beweis: Die Behauptung folgt schlicht aus der Tatsache, dass es sich hier ebenfalls um einen Tensorinterpolationsoperator handelt, so dass die einzelnen Interpolationsoperatoren umsortiert werden können. Die Kombination des Lemmas 4.28 und der Gleichung (4.3.3) liefert die Behauptung

$$\begin{aligned} &\left\| \mathfrak{I}^{c_L}_{Q_{t_L} \times Q_{s_L}} \circ \cdots \circ \mathfrak{I}^{c_1}_{Q_{t_1} \times Q_{s_1}} \right\|_{op, C(Q_{t_L} \times Q_{s_L}) \leftarrow C(Q_{t_0} \times Q_{s_0})} \leq \prod_{i=1}^{6} \left\| \mathfrak{I}_I^{i, \mathfrak{s}_c^L} \right\|_{op, C(I_L) \leftarrow C(I_0)} \\ &\leq \Lambda_m^6 (p^m c_{q,\varrho} \widehat{c}_{\gamma_\varrho, q, \varrho} + 1)^6. \end{aligned}$$

□

Bemerkung 27 (Kleine Cluster): *Unterschreiten die Cluster eine gewisse Größe in Relation zur Wellenzahl, existiert nur noch die Null-Richtung $c = 0$. Folglich tritt beim Wechsel zwischen solchen Clustern kein Interpolationsfehler auf. Praktisch bedeutet dies, dass* (2.3.3) *eingetreten ist. Die Tiefe des Pfades im Baum, entlang dem eine Reinterpolation nötig ist, steht dann in Relation zur Wellenzahl. Es gilt $L \in \mathcal{O}(\ln(\kappa))$, so dass die Bedingung $m \geq \mathcal{C}_{\varrho,q}(1 + \ln(L))$ vereinfacht werden kann. Es existiert dann ein $\widetilde{\mathcal{C}}$, so dass $m \geq \widetilde{\mathcal{C}} \ln(\ln(\kappa))$ ausreichend ist [3, Bem. 5.9].*

Die Fehleranalyse der Reinterpolation im Fall des Einfachschichtoperators von Börm und Melenk in [3] nutzt eine alternative Herangehensweise. Die beiden betrachten fast durchgängig nur Fehler auf reellen Intervallen, was zu einer ähnlichen Aussage zum Reinterpolationsfehler, aber einer kleineren Stabilitätskonstanten führt. Die schlechtere Stabilitätsaussage der hier gezeigten Vorgehensweise rührt daher, dass für die Abschätzung der Operatornorm die Bernstein-Ellipsen verlassen werden müssen, um auf die reellen Intervalle zurückzukehren, was zur Multiplikation mit der m-ten Potenz des Halbachsenparameters ϱ führt. Leider erweist sich der Ansatz von Börm und Melenk als schwierig anschlussfähig für den Doppelschichtoperator, weshalb in dieser Arbeit der oben vorgestellte Ansatz gewählt wurde, trotz seiner schlechteren Aussagen für die Stabilität.

4.3.2 Fehler auf Blöcken

Ziel dieses Abschnitts ist es, die Fehlerabschätzung der Interpolation von Matrixeinträgen auf Teilblöcke zu übertragen. Eine wichtige Rolle kommt dabei dem Fehler der Reinterpolation der Lagrange-Polynomen zu. Mit der Analyse des Reinterpolationsfehlers von Börm und Melenk ist es auch leicht möglich, den Fehler für die Reinterpolation der Lagrange-Polynome anzugeben [3, Kor. 5.8]. Die entsprechende Aussage soll hier ohne Beweis zitiert werden.

Lemma 4.30 (Interpolation der Lagrange-Polynome)
Seien für ein $L \in \underline{p_{\mathcal{I}}}$ eine Sequenz von Clustern $\mathfrak{s}_t^L$, die die Kontraktionseigenschaft (4.3.5) *für ein $\mathcal{C}_k \in (0,1)$ erfüllt, sowie eine Sequenz an Richtungen $\mathfrak{s}_c^L$ gegeben. Zusätzlich sei ein $\gamma \in \mathbb{R}_{>0}$ gegeben, so dass*

$$\kappa \operatorname{diam}(Q_{t_{\ell-1}}) \|c_{\ell-1} - c_\ell\|_2 \leq \gamma \qquad \text{für alle } \ell \in \underline{L}$$

erfüllt ist, und es existieren $\Lambda, \lambda \in \mathbb{R}_{\geq 1}$, so dass (2.1.2) *gilt. Weiterhin sei zu $q \in (\mathcal{C}_k, 1)$ ein $\tilde{q} \in (q,1)$ gegeben, dann existiert ein $\widetilde{m} \in \mathbb{N}$, so dass*

$$\left\| \ell_{t_0 c_0, \widehat{\mu}} - \ell_{\widehat{\mu}}^{\mathfrak{s}_t^L, \mathfrak{s}_c^L} \right\|_{\infty, Q_{t_L}} \leq \tilde{q}^m \qquad \text{für alle } m > \widetilde{m}, \widehat{\mu} \in \widehat{M}$$

gilt.

Die Aussage folgt analog für die Lagrange-Polynome der Spaltencluster, da komplexes Konjugieren auf den Betrag keinen Einfluss hat.
Mit diesem Resultat wiederum ist es mit dem Vorgehen wie in [4, S. 132 f.] möglich, den Fehler auf einem einzelnen Block abzuschätzen, auch wenn eine Reinterpolation für die Matrizen der Clusterbasis nötig ist.

Satz 4.31 (Fehler je Eintrag)
Gelten die Voraussetzungen vom Theorem 4.19 mit dem Zusatz $r = \min\left\{1, \frac{3}{4}\mathfrak{d}\right\}$ für Korollar 4.20 sowie von Lemma 4.30, dann existieren für einen zulässigen Block $b = (t,s,c) \in \mathcal{L}_{\mathcal{I}\times\mathcal{I}}^+$ mit der Interpolationsordnung $m > \widetilde{m}$ mit $\widetilde{m}$ aus Lemma 4.30 ein $\theta_{\tilde{q},\varrho} \in \mathbb{R}_{>0}$ mit $\theta_{\tilde{q},\varrho} < 1$ und ein $\zeta_{\eta_1,\eta_2}(b) \in \mathbb{R}_{>0}$, so dass

$$\left|(A_e)_{ij} - (\widetilde{A}_e)_{ij}\right| \leq \frac{e^{(\zeta_{\eta_1,\eta_2}(b) - \theta_{\tilde{q},\varrho})m}}{\operatorname{dist}(Q_t, Q_s)} \|\phi_i\|_{L^1(\Gamma)} \|\phi_j\|_{L^1(\Gamma)} \qquad \text{für alle } i \in {}^{\mathfrak{I}}t, j \in {}^{\mathfrak{I}}s$$

gilt, wobei ϕ_i, ϕ_j die reellwertigen Basisfunktionen sind.

Beweis: Sei ein zulässiger Block $b = (t,s,c) \in \mathcal{L}_{\mathcal{I}\times\mathcal{I}}^+$ gegeben, dann gibt es zwei mögliche Fälle zu unterscheiden.
1. Fall Im einfachsten Fall sind t, s Blätter, so dass keine Reinterpolation notwendig wird.

Für $i \in {}^{\mathfrak{I}}t$ und $j \in {}^{\mathfrak{I}}s$ sind die Matrixeinträge durch

$$\begin{aligned}(A_e)_{ij} &= \int_\Gamma \int_\Gamma \phi_i(x) g_e(x,y) \phi_j(y) \,\mathrm{d}x \,\mathrm{d}y \\ &= \int_\Gamma \int_\Gamma \phi_i(x) e^{\mathrm{i}\kappa\langle x-y,c\rangle_2} g_{ec}(x,y) \phi_j(y) \,\mathrm{d}x \,\mathrm{d}y,\end{aligned}$$

gegeben, während die approximierten Matrixeinträge durch

$$(\widetilde{A}_e)_{ij} = \int_\Gamma \int_\Gamma \phi_i(x) e^{\mathrm{i}\kappa\langle x-y,c\rangle_2} \mathfrak{I}_{Q_t \times Q_s}[g_{ec}](x,y) \phi_j(y) \,\mathrm{d}x \,\mathrm{d}y$$

berechnet werden. Aufgrund von (4.2.1) folgt für den Betrag der Differenz

$$\begin{aligned}\left|(A_e)_{ij} - (\widetilde{A}_e)_{ij}\right| &= \left| \int_\Gamma \int_\Gamma \phi_i(x) e^{\mathrm{i}\kappa\langle x-y,c\rangle_2} (g_{ec}(x,y) - \mathfrak{I}_{Q_t \times Q_s}[g_{ec}](x,y)) \phi_j(y) \,\mathrm{d}x \,\mathrm{d}y \right| \\ &\leq \int_\Gamma \int_\Gamma |\phi_i(x)| \, |g_{ec}(x,y) - \mathfrak{I}_{Q_t \times Q_s}[g_{ec}](x,y)| \, |\phi_j(y)| \,\mathrm{d}x \,\mathrm{d}y \\ &\leq \|g_{ec} - \mathfrak{I}_{Q_t \times Q_s}[g_{ec}]\|_{\infty, Q_t \times Q_s} \int_\Gamma \int_\Gamma |\phi_i(x)| \,\mathrm{d}x |\phi_j(y)| \,\mathrm{d}y \\ &= \|g_{ec} - \mathfrak{I}_{Q_t \times Q_s}[g_{ec}]\|_{\infty, Q_t \times Q_s} \, \|\phi_i\|_{L^1(\Gamma)} \|\phi_j\|_{L^1(\Gamma)},\end{aligned}$$

wobei die Norm des Interpolationsfehlers wie in Korollar 4.20 beschränkt ist. Die weitere Abschätzung, die zu der in der Behauptung angegebenen Schranke führt, folgt am Ende des zweiten Falls.

2. Fall Mindestens eines der Cluster t, s ist kein Blatt, so dass eine Reinterpolation notwendig wird. Fixiere Indizes $i \in {}^{\mathfrak{I}}t$ und $j \in {}^{\mathfrak{I}}s$. Seien durch $\mathfrak{s}_t^L$ und $\mathfrak{s}_s^L$ mit $L \in \underline{p_{\mathcal{I}|}}$ die Clustersequenzen gegeben, die von t, s über ihre Nachfahren bis zu den Blättern $t', s' \in \mathcal{L}_\mathcal{I}$ reichen, für die $i \in {}^{\mathfrak{I}}t', j \in {}^{\mathfrak{I}}s'$ gilt. Bezeichne durch $\mathfrak{s}_c^L$ die dazugehörige Sequenz an Richtungen. Die Approximation kann dann mit den reinterpolierten Lagrange-Polynomen (4.3.2) ausgeschrieben werden

$$(\widetilde{A}_e)_{ij} = \sum_{\widehat{\mu},\widehat{\nu} \in \widehat{M}} \int_\Gamma \int_\Gamma \phi_i(x) \ell^{\mathfrak{s}_t^L, \mathfrak{s}_c^L}_{\widehat{\mu}}(x) g_{ec}(x_{\widehat{\mu}}, y_{\widehat{\nu}}) \overline{\ell^{\mathfrak{s}_s^L, \mathfrak{s}_c^L}_{\widehat{\nu}}}(y) \phi_j(y) \,\mathrm{d}x \,\mathrm{d}y.$$

Um Aussagen über die zu untersuchende Differenz treffen zu können, füge die Interpolation in (t, s, c) als nahrhafte Null, also

$$\pm \sum_{\widehat{\mu},\widehat{\nu} \in \widehat{M}} \int_\Gamma \int_\Gamma \phi_i(x) \ell_{tc,\widehat{\mu}}(x) g_{ec}(x_{\widehat{\mu}}, y_{\widehat{\nu}}) \overline{\ell_{sc,\widehat{\nu}}}(y) \phi_j(y) \,\mathrm{d}x \,\mathrm{d}y,$$

ein, so dass nach dem Anwenden der Dreiecksungleichung zwei Beträge entstehen

$$\begin{aligned}&\left| \int_\Gamma \int_\Gamma \phi_i(x) e^{\mathrm{i}\kappa\langle x-y,c\rangle_2} (g_{ec}(x,y) - \mathfrak{I}_{Q_t \times Q_s}[g_{ec}](x,y)) \phi_j(y) \,\mathrm{d}x \,\mathrm{d}y \right| + \\ &\left| \sum_{\widehat{\mu},\widehat{\nu} \in \widehat{M}} \int_\Gamma \int_\Gamma \phi_i(x) \left(\ell_{tc,\widehat{\mu}}(x) \overline{\ell_{sc,\widehat{\nu}}}(y) - \ell^{\mathfrak{s}_t^L, \mathfrak{s}_c^L}_{\widehat{\mu}}(x) \overline{\ell^{\mathfrak{s}_s^L, \mathfrak{s}_c^L}_{\widehat{\nu}}}(y) \right) g_{ec}(x_{\widehat{\mu}}, y_{\widehat{\nu}}) \phi_j(y) \,\mathrm{d}x \,\mathrm{d}y \right|.\end{aligned}$$

Der erste Summand entspricht dem Fehler im ersten Fall und muss von daher nicht weiter untersucht werden. Im zweiten Betrag sind Aussagen zu Differenzen der verschiedenen Lagrange-Polynome notwendig, deren Abschätzung noch weitere Umformungen erfordert. Die Dreiecksungleichung liefert zunächst eine obere Schranke mit

$$\sum_{\widehat{\mu},\widehat{\nu}\in\widehat{M}} |g_{ec}(x_{\widehat{\mu}},y_{\widehat{\nu}})| \int_\Gamma\int_\Gamma |\phi_i(x)| \left|\ell_{tc,\widehat{\mu}}(x)\overline{\ell_{sc,\widehat{\nu}}}(y) - \ell_{\widehat{\mu}}^{\mathfrak{s}_t^L,\mathfrak{s}_c^L}(x)\overline{\ell_{\widehat{\nu}}^{\mathfrak{s}_s^L,\mathfrak{s}_c^L}}(y)\right| |\phi_j(y)| \,\mathrm{d}x\,\mathrm{d}y.$$

Der Betrag der Lagrange-Polynome kann durch Umschreiben in eine komplizierte Form auf Bekanntes zurückgeführt werden, dazu betrachte den Subtrahenden

$$\begin{aligned}&\ell_{\widehat{\mu}}^{\mathfrak{s}_t^L,\mathfrak{s}_c^L}(x)\overline{\ell_{\widehat{\nu}}^{\mathfrak{s}_s^L,\mathfrak{s}_c^L}}(y)\\ &= \left(\ell_{tc,\widehat{\mu}}(x) - \left(\ell_{tc,\widehat{\mu}}(x) - \ell_{\widehat{\mu}}^{\mathfrak{s}_t^L,\mathfrak{s}_c^L}(x)\right)\right)\left(\overline{\ell_{sc,\widehat{\nu}}}(y) - \left(\overline{\ell_{sc,\widehat{\nu}}}(y) - \overline{\ell_{\widehat{\nu}}^{\mathfrak{s}_s^L,\mathfrak{s}_c^L}}(y)\right)\right).\end{aligned}$$

Ausmultiplizieren zusammen mit der Dreiecksungleichung liefert für die Subtraktion

$$\begin{aligned}&\left|\ell_{tc,\widehat{\mu}}(x)\overline{\ell_{sc,\widehat{\nu}}}(y) - \ell_{\widehat{\mu}}^{\mathfrak{s}_t^L,\mathfrak{s}_c^L}(x)\overline{\ell_{\widehat{\nu}}^{\mathfrak{s}_s^L,\mathfrak{s}_c^L}}(y)\right|\\ &\quad\overset{\Delta}{\leq} \left|\ell_{tc,\widehat{\mu}}(x)\right|\left|\overline{\ell_{sc,\widehat{\nu}}}(y) - \overline{\ell_{\widehat{\nu}}^{\mathfrak{s}_s^L,\mathfrak{s}_c^L}}(y)\right| + \left|\ell_{tc,\widehat{\mu}}(x) - \ell_{\widehat{\mu}}^{\mathfrak{s}_t^L,\mathfrak{s}_c^L}(x)\right|\left|\overline{\ell_{sc,\widehat{\nu}}}(y)\right|\\ &\qquad + \left|\ell_{tc,\widehat{\mu}}(x) - \ell_{\widehat{\mu}}^{\mathfrak{s}_t^L,\mathfrak{s}_c^L}(x)\right|\left|\overline{\ell_{sc,\widehat{\nu}}}(y) - \overline{\ell_{\widehat{\nu}}^{\mathfrak{s}_s^L,\mathfrak{s}_c^L}}(y)\right|.\end{aligned}$$

Die Beträge der einzelnen Lagrange-Polynome können gegen die Stabilitätskonstante 2.3 der Interpolation im Dreidimensionalen abgeschätzt werden, so dass nur Fehler vom Typ des Lemmas 4.30 übrig bleiben

$$\begin{aligned}\left|\ell_{tc,\widehat{\mu}}(x)\overline{\ell_{sc,\widehat{\nu}}}(y) - \ell_{\widehat{\mu}}^{\mathfrak{s}_t^L,\mathfrak{s}_c^L}(x)\overline{\ell_{\widehat{\nu}}^{\mathfrak{s}_s^L,\mathfrak{s}_c^L}}(y)\right| &\leq \Lambda_m^3\tilde{q}^m + \tilde{q}^m\Lambda_m^3 + \tilde{q}^m\tilde{q}^m\\ &= \tilde{q}^m(2\Lambda_m^3 + \tilde{q}^m).\end{aligned}$$

Die Funktionsauswertungen von g_{ec} können mit der Eigenschaft der ebenen Welle (4.2.1) in Kombination mit $x_{\widehat{\mu}} \in Q_t$ sowie $y_{\widehat{\nu}} \in Q_s$ für alle $\widehat{\mu},\widehat{\nu} \in \widehat{M}$ beschränkt werden

$$|g_{ec}(x_{\widehat{\mu}},y_{\widehat{\nu}})| = \frac{\left|e^{\mathrm{i}\kappa\left(\|x_{\widehat{\mu}}-y_{\widehat{\nu}}\|_2 - \langle x_{\widehat{\mu}}-y_{\widehat{\nu}},c\rangle_2\right)}\right|}{4\pi\|x_{\widehat{\mu}}-y_{\widehat{\nu}}\|_2} \leq \frac{1}{4\pi\,\mathrm{dist}(Q_t,Q_s)}.$$

Da aus der Tensorinterpolation $\#\widehat{M} = (m+1)^3$ folgt, ergibt sich

$$\begin{aligned}&\sum_{\widehat{\mu},\widehat{\nu}\in\widehat{M}} |g_{ec}(x_{\widehat{\mu}},y_{\widehat{\nu}})| \int_\Gamma\int_\Gamma |\phi_i(x)| \left|\ell_{tc,\widehat{\mu}}(x)\overline{\ell_{sc,\widehat{\nu}}}(y) - \ell_{\widehat{\mu}}^{\mathfrak{s}_t^L,\mathfrak{s}_c^L}(x)\overline{\ell_{\widehat{\nu}}^{\mathfrak{s}_s^L,\mathfrak{s}_c^L}}(y)\right| |\phi_j(y)| \,\mathrm{d}x\,\mathrm{d}y\\ &\leq \sum_{\widehat{\mu},\widehat{\nu}\in\widehat{M}} \frac{1}{4\pi\,\mathrm{dist}(Q_t,Q_s)} \int_\Gamma\int_\Gamma |\phi_i(x)|\tilde{q}^m(2\Lambda_m^3+\tilde{q}^m)|\phi_j(y)| \,\mathrm{d}x\,\mathrm{d}y\\ &\leq \frac{(m+1)^6}{4\pi\,\mathrm{dist}(Q_t,Q_s)}\tilde{q}^m(2\Lambda_m^3+\tilde{q}^m)\|\phi_i\|_{L^1(\Gamma)}\|\phi_j\|_{L^1(\Gamma)}.\end{aligned}$$

Damit kann der zweite Fall insgesamt gegen

$$\begin{aligned}
&\left(\|g_{ec} - \mathfrak{I}_{Q_t\times Q_s}[g_{ec}]\|_{\infty,Q_t\times Q_s} + \tfrac{(m+1)^6}{4\pi\,\mathrm{dist}(Q_t,Q_s)}\tilde{q}^m(2\Lambda_m^3+\tilde{q}^m)\right)\|\phi_i\|_{L^1(\Gamma)}\|\phi_j\|_{L^1(\Gamma)}\\
&\leq \left((1+\Lambda_m)12\Lambda_m^5\varrho^{-m}\tfrac{\mathcal{C}_{\eta_2}e^{\eta_1+\eta_2}}{\pi\,\mathrm{dist}(Q_t,Q_s)} + \tfrac{(m+1)^6}{4\pi\,\mathrm{dist}(Q_t,Q_s)}\tilde{q}^m(2\Lambda_m^3+\tilde{q}^m)\right)\|\phi_i\|_{L^1(\Gamma)}\|\phi_j\|_{L^1(\Gamma)}\\
&\leq \tfrac{\|\phi_i\|_{L^1(\Gamma)}\|\phi_j\|_{L^1(\Gamma)}}{\mathrm{dist}(Q_t,Q_s)}\left((1+\Lambda_m)12\Lambda_m^5\varrho^{-m}\tfrac{\mathcal{C}_{\eta_2}e^{\eta_1+\eta_2}}{\pi} + \tfrac{(m+1)^6}{4\pi}\tilde{q}^m(2\Lambda_m^3+\tilde{q}^m)\right)
\end{aligned}$$

abgeschätzt werden. Unter Berücksichtigung von $\tilde{q}<1$ sowie $\varrho>1$ setze $\theta := \max\left\{\tilde{q},\frac{1}{\varrho}\right\}$ und $\theta_{\tilde{q},\varrho} = 1-\theta$. Damit gilt $\theta_{\tilde{q},\varrho} = \min\left\{1-\tilde{q},1-\frac{1}{\varrho}\right\}$, so dass

$$\begin{aligned}
&(1+\Lambda_m)12\Lambda_m^5\varrho^{-m}\tfrac{\mathcal{C}_{\eta_2}e^{\eta_1+\eta_2}}{\pi} + \tfrac{(m+1)^6}{4\pi}\tilde{q}^m(2\Lambda_m^3+\tilde{q}^m)\\
&\leq (\theta)^m\left((1+\Lambda_m)12\Lambda_m^5\tfrac{\mathcal{C}_{\eta_2}e^{\eta_1+\eta_2}}{\pi} + \tfrac{(m+1)^6}{4\pi}(2\Lambda_m^3+1)\right)\\
&= (1-\theta_{\tilde{q},\varrho})^m\left((1+\Lambda_m)12\Lambda_m^5\tfrac{\mathcal{C}_{\eta_2}e^{\eta_1+\eta_2}}{\pi} + \tfrac{(m+1)^6}{4\pi}(2\Lambda_m^3+1)\right)
\end{aligned}$$

folgt. Da die Exponentialfunktion auf ganz $\mathbb{R}$ stetig differenzierbar ist, existiert nach dem Mittelwertsatz der Differentialrechnung zu jedem $\theta_{\tilde{q},\varrho}\in(0,1)$ ein $\zeta\in(-1,0)$ mit

$$e^{\zeta} = \tfrac{e^0-e^{-\theta_{\tilde{q},\varrho}}}{0-(-\theta_{\tilde{q},\varrho})} \qquad\Longleftrightarrow\qquad e^{\zeta}\theta_{\tilde{q},\varrho} = 1-e^{-\theta_{\tilde{q},\varrho}}$$

und da $e^{\zeta}<1$ gilt, folgt

$$\theta_{\tilde{q},\varrho} > e^{\zeta}\theta_{\tilde{q},\varrho} = 1-e^{-\theta_{\tilde{q},\varrho}} \qquad\Longrightarrow\qquad e^{-\theta_{\tilde{q},\varrho}} > 1-\theta_{\tilde{q},\varrho}.$$

Dies kann genutzt werden, um

$$(1-\theta_{\tilde{q},\varrho})^m \leq (e^{-\theta_{\tilde{q},\varrho}})^m = e^{-m\theta_{\tilde{q},\varrho}}$$

zu erhalten. Mit der Bemerkung 6 kann der Term

$$(1+\Lambda_m)12\Lambda_m^5\tfrac{\mathcal{C}_{\eta_2}e^{\eta_1+\eta_2}}{\pi} + \tfrac{(m+1)^6}{4\pi}(2\Lambda_m^3+1)$$

für feste η_1 und η_2 durch ein Polynom mit Variable m abgeschätzt werden, demnach existiert ein von Λ und λ abhängiges minimales $\zeta_{\eta_1,\eta_2}(b)\in\mathbb{R}_{>0}$ mit

$$(1+\Lambda_m)12\Lambda_m^5\tfrac{\mathcal{C}_{\eta_2}e^{\eta_1+\eta_2}}{\pi} + \tfrac{(m+1)^6}{4\pi}(2\Lambda_m^3+1) \leq e^{m\zeta_{\eta_1,\eta_2}(b)}.$$

Dabei gilt, dass für wachsendes m immer kleinere Werte von ζ_{η_1,η_2} ausreichen. Insgesamt kann die Abschätzung

$$\begin{aligned}
&\left(\|g_{ec} - \mathfrak{I}_{Q_t\times Q_s}[g_{ec}]\|_{\infty,Q_t\times Q_s} + \tfrac{(m+1)^6}{4\pi\,\mathrm{dist}(Q_t,Q_s)}\tilde{q}^m(2\Lambda_m^3+\tilde{q}^m)\right)\|\phi_i\|_{L^1(\Gamma)}\|\phi_j\|_{L^1(\Gamma)}\\
&\leq \tfrac{e^{(\zeta_{\eta_1,\eta_2}(b)-\theta_{\tilde{q},\varrho})m}}{\mathrm{dist}(Q_t,Q_s)}\|\phi_i\|_{L^1(\Gamma)}\|\phi_j\|_{L^1(\Gamma)}
\end{aligned}$$

gemacht werden. Sowohl der erste als auch der zweite Fall lassen sich durch diesen Term beschränken.
Wenn die Interpolationsordnung m hoch genug ist und damit $\zeta_{\eta_1,\eta_2}(b) < \theta_{\tilde{q},\varrho}$ erfüllt wird, zeigt sich die zu erwartende exponentielle Konvergenz. □

Bemerkung 28 (Konvergenz des Fehlers je Eintrag): *Wann die für die exponentielle Konvergenz auf dem Block b notwendige Bedingung $\zeta_{\eta_1,\eta_2}(b) < \theta_{\tilde{q},\varrho}$ erreicht ist, hängt einerseits von $\widetilde{m}$ und damit von $L, \mathcal{C}_k, \gamma$ und dem gewählten q aus Lemma 4.30 sowie den Konstanten Λ, λ des verwendeten Interpolationsansatzes, den Zulässigkeitsparametern η_1, η_2, andererseits durch $\theta_{\tilde{q},\varrho}$ auch von $\tilde{q}$ und dem Halbachsenparameter ϱ der gewählten Bernstein-Ellipse ab.*
Entsprechend gilt es, mit den konkreten Parametern das nötige $m \in \mathbb{N}_{>\widetilde{m}}$ mit

$$(1-\theta_{\tilde{q},\varrho})^m \left((1+\Lambda_m)12\Lambda_m^5 \tfrac{\mathcal{C}_{\eta_2} e^{\eta_1+\eta_2}}{\pi} + \tfrac{(m+1)^6}{4\pi}(2\Lambda_m^3+1)\right) < e^{m\mathcal{C}}$$

für ein $\mathcal{C} < 1$ zu bestimmen.

Zu guter Letzt sollen mit den bisherigen Resultaten Aussagen zu der Genauigkeit einzelner Matrixblöcke geschaffen werden, die dann auf die gesamte Matrix übertragen werden können. Dazu sind Matrixnormen nötig.

Definition 4.32 (Matrixnorm)
Seien $\mathcal{I}, \mathcal{J} \subset \mathbb{N}$ gegeben. Eine Abbildung $\|\cdot\| : \mathbb{C}^{\mathcal{I}\times\mathcal{J}} \to \mathbb{R}_{\geq 0}$ heißt Matrixnorm, falls sie die drei Axiome einer Norm erfüllt.
Durch eine Vektornorm $\|\cdot\|_V$ wird über

$$\|A\| := max\left\{\tfrac{\|Ax\|_V}{\|x\|_V} \mid x \in \mathbb{C}^{\mathcal{J}} \setminus \{0\}\right\} \qquad \text{für alle } A \in \mathbb{C}^{\mathcal{I}\times\mathcal{J}},$$

beziehungsweise über die äquivalente Formulierung

$$\|A\| = max\left\{\|Ax\|_V \mid x \in \mathbb{C}^{\mathcal{J}}, \|x\|_V = 1\right\} \qquad \text{für alle } A \in \mathbb{C}^{\mathcal{I}\times\mathcal{J}}.$$

eine Matrixnorm definiert, sie heißt auch induzierte Matrixnorm.

Von der euklidischen Norm wird die Spektralnorm für Matrizen induziert.

Definition und Lemma 4.33 (Spektralnorm)
Für $\mathcal{I}, \mathcal{J} \subset \mathbb{N}$ ist die Spektralnorm $\|\cdot\|_2$ einer Matrix $A \in \mathbb{C}^{\mathcal{I}\times\mathcal{J}}$ durch

$$\|A\|_2 := max\left\{\tfrac{\|Ax\|_2}{\|x\|_2} \mid x \in \mathbb{C}^{\mathcal{J}} \setminus \{0\}\right\}$$

gegeben. Die Spektralnorm erfüllt $\|A\|_2 = \|A^*\|_2$ und ist mit der euklidischen Norm verträglich, das heißt, für alle Vektoren $x \in \mathbb{C}^{\mathcal{J}}$ gilt

$$\|Ax\|_2 \leq \|A\|_2\|x\|_2.$$

Oftmals ist es sehr hilfreich, wenn Abschätzungen für hierarchische Matrizen zunächst in der *Frobeniusnorm*[vii] gemacht werden, auch wenn eine Aussage in der Spektralnorm gewünscht ist.

Definition und Lemma 4.34 (Frobeniusnorm)
Die Frobeniusnorm einer Matrix $A \in \mathbb{C}^{\mathcal{I}\times\mathcal{J}}$ für $\mathcal{I}, \mathcal{J} \subset \mathbb{N}$ ist durch

$$\|A\|_F := \left(\sum_{i\in\mathcal{I}} \sum_{j\in\mathcal{J}} |a_{ij}|^2 \right)^{\frac{1}{2}}$$

gegeben und erfüllt ebenfalls $\|A\|_F = \|A^*\|_F$. Die Frobeniusnorm kann auch über das Frobeniusskalarprodukt dargestellt werden $\|A\|_F^2 = \langle A, A\rangle_F$.

Die Frobeniusnorm bietet den Vorteil, dass sie exakt durch die Norm von Teilmatrizen berechnet werden kann. Für die Untersuchung des Fehlers bedeutet dies, dass jede Teilmatrix unabhängig von den anderen analysiert werden kann und eine Fallunterscheidung für zulässige und unzulässige Teilmatrizen so leicht zugänglich ist. Jedoch handelt es sich bei der Frobeniusnorm um keine induzierte Matrixnorm, sie ist aber mit der euklidischen Norm verträglich.

Lemma 4.35 (Globale Frobeniusnorm)
Seien ein richtungsabhängiger Blockbaum $T_{\mathcal{I}\times\mathcal{I}}$ und eine Matrix $A \in \mathbb{C}^{\mathcal{I}\times\mathcal{I}}$ gegeben. Die globale Frobeniusnorm der Matrix A lässt sich blockweise durch die Frobeniusnorm darstellen, es gilt

$$\|A\|_F = \left(\sum_{(t,s,c)\in\mathcal{L}_{\mathcal{I}\times\mathcal{I}}} \|A|_{\star(t\times s)}\|_F^2 \right)^{\frac{1}{2}}.$$

Beweis: Die Behauptung ergibt sich direkt durch Nachrechnen oder kann auch in [4, Lem. 4.39] nachgeschlagen werden. □

Die Spektralnorm einer Matrix kann immer durch ihre Frobeniusnorm beschränkt werden, so dass mit der Untersuchung der Frobeniusnorm glücklicherweise auch immer eine Abschätzung der Spektralnorm gegeben ist.

Lemma 4.36 (Abgeschätzte Spektralnorm)
Für jede Matrix $A \in \mathbb{C}^{\mathcal{I}\times\mathcal{I}}$ gilt

$$\|A\|_2 \leq \|A\|_F.$$

[vii] Benannt nach dem deutschen Mathematiker Ferdinand Georg Frobenius.

Beweis: Ein Beweis befindet sich in [4, S. 120 f.] als Teil des Beweises von Lemma 4.44.

Der Satz 4.31 für den Fehler je Eintrag erlaubt dann eine Fehlerabschätzung in der Frobeniusnorm auf dem gesamten Block b nach dem typischen Vorgehen wie in [4].

Satz 4.37 (Fehler auf Blöcken)
Seien die Voraussetzung von Satz 4.31 erfüllt und die gewählten Basisfunktionen quadratintegrierbar, dann kann der Fehler eines zulässigen Blocks $b = (t, s, c) \in \mathcal{L}^+_{\mathcal{I}\times\mathcal{I}}$ in der Frobeniusnorm durch

$$\left\|A_e|_{t\times s} - \widetilde{A}_e|_{t\times s}\right\|_F \leq \mathcal{C}_{L^2(\Gamma),t} \quad \mathcal{C}_{L^2(\Gamma),s} \quad \frac{e^{(\zeta_{\eta_1,\eta_2}(b)-\theta_{\tilde{q},\varrho})m}}{\operatorname{dist}(Q_t,Q_s)} \left(\sum_{i\in{}^{\mathfrak{I}}t} |\Gamma_i|\right)^{\frac{1}{2}} \left(\sum_{j\in{}^{\mathfrak{I}}s} |\Gamma_j|\right)^{\frac{1}{2}}$$

mit den von der Triangulation abhängigen Konstanten

$$\mathcal{C}_{L^2(\Gamma),t} := \max\left\{\|\phi_i\|_{L^2(\Gamma)} \mid i \in {}^{\mathfrak{I}}t\right\}, \quad \mathcal{C}_{L^2(\Gamma),s} := \max\left\{\|\phi_j\|_{L^2(\Gamma)} \mid j \in {}^{\mathfrak{I}}s\right\}$$

beschränkt werden.

Beweis: Für jedes $i \in {}^{\mathfrak{I}}t$, $j \in {}^{\mathfrak{I}}s$ gilt die Abschätzung aus Satz 4.31, so dass mit der Definition der Frobeniusnorm direkt

$$\begin{aligned}
\left\|A_e|_{t\times s} - \widetilde{A}_e|_{t\times s}\right\|_F^2 &= \sum_{i\in{}^{\mathfrak{I}}t}\sum_{j\in{}^{\mathfrak{I}}s} \left|(A_e)_{ij} - (\widetilde{A}_e)_{ij}\right|^2 \\
&\leq \sum_{i\in{}^{\mathfrak{I}}t}\sum_{j\in{}^{\mathfrak{I}}s} \left(\frac{e^{(\zeta_{\eta_1,\eta_2}(b)-\theta_{\tilde{q},\varrho})m}}{\operatorname{dist}(Q_t,Q_s)} \|\phi_i\|_{L^1(\Gamma)}\|\phi_j\|_{L^1(\Gamma)}\right)^2 \\
&= \left(\frac{e^{(\zeta_{\eta_1,\eta_2}(b)-\theta_{\tilde{q},\varrho})m}}{\operatorname{dist}(Q_t,Q_s)}\right)^2 \sum_{i\in{}^{\mathfrak{I}}t} \|\phi_i\|^2_{L^1(\Gamma)} \sum_{j\in{}^{\mathfrak{I}}s} \|\phi_j\|^2_{L^1(\Gamma)}
\end{aligned}$$

folgt. Die L^1-Norm kann mit der Cauchy-Schwarz-Ungleichung gegen die L^2-Norm abgeschätzt werden. Um den zusätzlichen Term möglichst klein zu halten, wird das Integral zunächst auf den Träger $\Gamma_i \subset \Gamma$ der Basisfunktion ϕ_i reduziert

$$\begin{aligned}
\sum_{i\in{}^{\mathfrak{I}}t} \|\phi_i\|^2_{L^1(\Gamma)} &= \sum_{i\in{}^{\mathfrak{I}}t} \left|\int_\Gamma |\phi_i(x)|\,\mathrm{d}x\right|^2 = \sum_{i\in{}^{\mathfrak{I}}t} \left|\int_{\Gamma_i} |\phi_i(x)|\,\mathrm{d}x\right|^2 = \sum_{i\in{}^{\mathfrak{I}}t} \left|\int_\Gamma |\phi_i(x)1_{\Gamma_i}|\,\mathrm{d}x\right|^2 \\
&\leq \sum_{i\in{}^{\mathfrak{I}}t} \int_\Gamma |1_{\Gamma_i}|^2\,\mathrm{d}x \int_\Gamma |\phi_i(x)|^2\,\mathrm{d}x = \sum_{i\in{}^{\mathfrak{I}}t} |\Gamma_i|\|\phi_i\|^2_{L^2(\Gamma)} \\
&\leq \mathcal{C}^2_{L^2(\Gamma),t} \sum_{i\in{}^{\mathfrak{I}}t} |\Gamma_i|.
\end{aligned}$$

Durch Wurzelziehen ergibt sich die gewünschte Abschätzung

$$\left\|A_e|_{t\times s} - \widetilde{A}_e|_{t\times s}\right\|_F \leq \mathcal{C}_{L^2(\Gamma),t}\mathcal{C}_{L^2(\Gamma),s} \frac{e^{(\zeta_{\eta_1,\eta_2}(b)-\theta_{\tilde{q},\varrho})m}}{\operatorname{dist}(Q_t,Q_s)} \left(\sum_{i\in{}^{\mathfrak{I}}t} |\Gamma_i|\right)^{\frac{1}{2}} \left(\sum_{j\in{}^{\mathfrak{I}}s} |\Gamma_j|\right)^{\frac{1}{2}}.$$

□

Mit zusätzlichen Voraussetzungen an die betrachtete Triangulation der Oberfläche kann die L^2-Norm der Basisfunktionen abgeschätzt werden [32, S. 124, 268 f.].
Wenn die Triangulation *formregulär* ist, also wenn der minimal auftretende Innenwinkel der Dreiecke nach unten beschränkt werden kann, existiert eine Konstante $\mathcal{C}_{To} \in \mathbb{N}$, die die maximale Anzahl an überlappenden Trägern beschränkt, es gilt

$$\sum_{i \in \mathcal{I}} |\Gamma_i| \leq \mathcal{C}_{To} |\Gamma|. \tag{4.3.13}$$

So beschränkt Formregularität zum Beispiel bei der Verwendung von stückweise linearen Basisfunktionen die Anzahl an Dreiecken der Triangulation, die sich einen Eckpunkt teilen.
Ist die betrachtete Oberflächentriangulation zusätzlich noch *quasi-uniform*, das heißt, weisen alle Dreiecke eine vergleichbare Größe auf, dann existiert eine Konstante $\mathcal{C}_{Th} \in \mathbb{R}_{>0}$ mit

$$\|\phi_i\|_{L^2(\Gamma)} \leq \mathcal{C}_{Th} h \qquad \text{für alle } i \in \mathcal{I}, \tag{4.3.14}$$

wobei h die maximal auftretende Kantenlänge der Dreiecke der Triangulation ist.

Korollar 4.38 (Fehler auf Blöcken)
Seien die Voraussetzung von Satz 4.31 und zusätzlich (4.3.14) *erfüllt sowie die gewählten Basisfunktionen quadratintegrierbar, dann kann der Fehler aus Satz 4.37 auf dem Block $b = (t, s, c) \in \mathcal{L}_{\mathcal{I}\times\mathcal{I}}$ durch*

$$\left\| A_e|_{t\times s} - \widetilde{A}_e|_{t\times s} \right\|_F \leq \mathcal{C}_{Th}^2 h^2 \frac{e^{(\zeta_{\eta_1,\eta_2}(b)-\theta_{\tilde{q},\varrho})m}}{\operatorname{dist}(Q_t,Q_s)} \left(\sum_{i \in \mathfrak{I}t} |\Gamma_i| \right)^{\frac{1}{2}} \left(\sum_{j \in \mathfrak{I}s} |\Gamma_j| \right)^{\frac{1}{2}}$$

beschränkt werden, wobei h die maximal auftretende Kantenlänge der Dreiecke der Triangulation ist.

Beweis: Mit der Voraussetzung (4.3.14) kann $\mathcal{C}_{L^2(\Gamma),t}$ durch

$$\mathcal{C}_{L^2(\Gamma),t} = \max \left\{ \|\phi_i\|_{L^2(\Gamma)} \mid i \in {}^{\mathfrak{I}}t \right\} \leq \mathcal{C}_{Th} h$$

abgeschätzt werden ebenso wie $\mathcal{C}_{L^2(\Gamma),s}$. □

Weiterhin kann ausgenutzt werden, dass die Frobeniusnorm sich additiv aus der Norm der Teilmatrizen zusammen setzt (siehe Lemma 4.35), um eine Aussage in der Gestalt von [4, Lem. 4.40] zu erhalten.
Aus der Quasi-Uniformität zusammen mit der Formregularität der Triangulation folgt auch,

dass die Differenz der minimalen und maximalen Kantenlänge beschränkt ist. Da die Distanz der Cluster eines zulässigen Clusterpaars größer als null sein muss und somit in einem Verhältnis zur minimalen Kantenlänge der Triangulation steht, existiert eine Konstante $\mathcal{C}_{Td} \in \mathbb{R}_{>0}$ mit

$$\min\left\{\operatorname{dist}(Q_t, Q_s) \mid (t,s,c) \in \mathcal{L}^+_{\mathcal{I}\times\mathcal{I}}\right\} > \mathcal{C}_{Td} h. \tag{4.3.15}$$

Theorem 4.39 (Gesamtfehler)
Der globale Approximationsfehler kann unter den Voraussetzungen von Satz 4.31 und der Annahme, dass die Triangulation quasi-uniform und formregulär ist und damit eine Überlappungskonstante $\mathcal{C}_{To} \in \mathbb{N}$ (4.3.13) existiert sowie dass die gewählten Basisfunktionen quadratintegrierbar sind, mit

$$\|A_e - \widetilde{A}_e\|_F \leq \mathcal{C}_{To}\mathcal{C}^2_{Th}\mathcal{C}^{-1}_{Td}|\Gamma| h \epsilon_{b,m}$$

abgeschätzt werden, wobei

$$\epsilon_{b,m} := \max\left\{ e^{(\zeta_{\eta_1,\eta_2}(b) - \theta_{\tilde{q},\varrho})m} \mid b = (t,s,c) \in \mathcal{L}^+_{\mathcal{I}\times\mathcal{I}} \right\},$$

$\mathcal{C}_{Th} \in \mathbb{R}_{>0}$ nach (4.3.14)*, $\mathcal{C}_{Td} \in \mathbb{R}_{>0}$ nach* (4.3.15) *und h die maximale Kantenlänge der Triangulation seien.*

Beweis: Für den Fehler der Gesamtmatrix gilt mit dem Lemma zur Blattpartition 2.29

$$\begin{aligned}
\|A_e - \widetilde{A}_e\|_F &= \left(\sum_{(t,s,c)\in\mathcal{L}_{\mathcal{I}\times\mathcal{I}}} \left\| A_e|_{t\times s} - \widetilde{A}_e|_{t\times s} \right\|_F^2 \right)^{\frac{1}{2}} \\
&= \left(\sum_{(t,s,c)\in\mathcal{L}^+_{\mathcal{I}\times\mathcal{I}}} \left\| A_e|_{t\times s} - \widetilde{A}_e|_{t\times s} \right\|_F^2 \right)^{\frac{1}{2}} \\
&\overset{4.38}{\leq} \left(\sum_{(t,s,c)\in\mathcal{L}^+_{\mathcal{I}\times\mathcal{I}}} \left(\mathcal{C}^2_{Th} h^2 \tfrac{e^{(\zeta_{\eta_1,\eta_2}(b) - \theta_{\tilde{q},\varrho})m}}{\operatorname{dist}(Q_t,Q_s)} \right)^2 \left(\sum_{i\in\mathfrak{I}t} |\Gamma_i| \right) \left(\sum_{j\in\mathfrak{I}s} |\Gamma_j| \right) \right)^{\frac{1}{2}} \\
&\leq \mathcal{C}^2_{Th} h^2 \tfrac{\epsilon_{b,m}}{\mathcal{C}_{Td} h} \left(\sum_{(t,s,c)\in\mathcal{L}^+_{\mathcal{I}\times\mathcal{I}}} \left(\sum_{i\in\mathfrak{I}t} |\Gamma_i| \right) \left(\sum_{j\in\mathfrak{I}s} |\Gamma_j| \right) \right)^{\frac{1}{2}} \\
&\leq \mathcal{C}^2_{Th}\mathcal{C}^{-1}_{Td} h \epsilon_{b,m} \left(\sum_{i\in\mathcal{I}} |\Gamma_i| \sum_{j\in\mathcal{I}} |\Gamma_j| \right)^{\frac{1}{2}} \\
&\leq \mathcal{C}^2_{Th}\mathcal{C}^{-1}_{Td} h \epsilon_{b,m} (\mathcal{C}^2_{To}|\Gamma|^2)^{\frac{1}{2}}.
\end{aligned}$$

□

Da die Frobeniusnorm eine obere Schranke für die Spektralnorm bildet (siehe Lemma 4.36), ist auf diese Weise auch eine Schranke für die Spektralnorm gefunden.

4.4 Doppelschichtoperator

Die Fehleranalyse des Doppelschichtoperators gestaltet sich etwas schwieriger. Die Funktion wird zunächst interpoliert und im Anschluss daran die Normalenableitung des Interpolationspolynoms gebildet und als Approximation verwendet.
Weil beim Differenzieren der Polynomgrad um eins reduziert wird und der Polynomgrad in direkter Relation zum Approximationsfehler steht, ist ein schlechterer Faktor in den Konvergenzaussagen zu erwarten, welcher sich durch die Durchmesserquadrate statt der Distanz als Divisor im Fehler und größeren Konstanten bemerkbar macht (siehe dazu Theorem 4.44).
Der äußere Normalenvektor existiert für fast alle Punkte $(x, y) \in \Gamma$, so dass für den Interpolationsfehler beim Doppelschichtoperator

$$\begin{aligned}\tfrac{\partial}{\partial n(y)}\left(g_e(x,y) - \mathfrak{I}^c_{Q_t\times Q_s}[g_e](x,y)\right) &= \tfrac{\partial}{\partial n(y)}\left(e^{\mathrm{i}\kappa\langle x-y,c\rangle_2}\left(g_{ec}(x,y) - \mathfrak{I}_{Q_t\times Q_s}[g_{ec}](x,y)\right)\right)\\ &= \langle \nabla_y\left(e^{\mathrm{i}\kappa\langle x-y,c\rangle_2}\left(g_{ec}(x,y) - \mathfrak{I}_{Q_t\times Q_s}[g_{ec}](x,y)\right)\right), n(y)\rangle_{\mathbb{C}}\end{aligned}$$

gilt, wobei $n(y)$ den Normaleneinheitsvektor im Punkt $y \in \Gamma$ bezeichne.
Anwenden der Cauchy-Schwarz-Ungleichung auf die Norm des Fehlers führt zu

$$\begin{aligned}&\max\left\{\left|\langle \nabla_y\left(e^{\mathrm{i}\kappa\langle x-y,c\rangle_2}\left(g_{ec}(x,y) - \mathfrak{I}_{Q_t\times Q_s}[g_{ec}](x,y)\right)\right), n(y)\rangle_{\mathbb{C}}\right| \mid (x,y) \in Q_t\times Q_s\right\}\\ &\leq \max\left\{\left\|\nabla_y\left(e^{\mathrm{i}\kappa\langle x-y,c\rangle_2}\left(g_{ec}(x,y) - \mathfrak{I}_{Q_t\times Q_s}[g_{ec}](x,y)\right)\right)\right\|_2 \mid (x,y) \in Q_t\times Q_s\right\}.\end{aligned}$$

Die Norm $\|n(y)\|_2$ kann weggelassen werden, da die Länge des Normaleneinheitsvektors eins beträgt. Die Verwendung der Produkt- und Kettenregel führt zu dem Ausdruck

$$\begin{aligned}&\nabla_y\left(e^{\mathrm{i}\kappa\langle x-y,c\rangle_2}\left(g_{ec}(x,y) - \mathfrak{I}_{Q_t\times Q_s}[g_{ec}](x,y)\right)\right) = e^{\mathrm{i}\kappa\langle x-y,c\rangle_2}\\ &\cdot\left(\nabla_y\Big(g_{ec}(x,y) - \mathfrak{I}_{Q_t\times Q_s}[g_{ec}](x,y)\Big) - \mathrm{i}\kappa c\Big(g_{ec}(x,y) - \mathfrak{I}_{Q_t\times Q_s}[g_{ec}](x,y)\Big)\right).\end{aligned}$$

Um die euklidische Norm des letzten Ausdrucks abschätzen zu können, nutze die Dreiecksungleichung und das i, c sowie die ebene Welle jeweils von Norm Eins sind

$$\begin{aligned}&\left\|\nabla_y\left(e^{\mathrm{i}\kappa\langle x-y,c\rangle_2}\left(g_{ec}(x,y) - \mathfrak{I}_{Q_t\times Q_s}[g_{ec}](x,y)\right)\right)\right\|_2\\ &\quad= \left\|\nabla_y\left(g_{ec}(x,y) - \mathfrak{I}_{Q_t\times Q_s}[g_{ec}](x,y)\right) - \mathrm{i}\kappa c\left(g_{ec}(x,y) - \mathfrak{I}_{Q_t\times Q_s}[g_{ec}](x,y)\right)\right\|_2\\ &\quad\overset{\Delta}{\leq} \left\|\nabla_y\left(g_{ec}(x,y) - \mathfrak{I}_{Q_t\times Q_s}[g_{ec}](x,y)\right)\right\|_2 + \kappa\left|g_{ec}(x,y) - \mathfrak{I}_{Q_t\times Q_s}[g_{ec}](x,y)\right|.\end{aligned}$$

Zusätzlich kann die Normäquivalenz der euklidischen Norm und der Maximum-Norm genutzt werden, um erneut für den ersten Summanden eine Abschätzung in der Maximum-Norm zu erhalten

$$\begin{aligned}&\max\left\{\|\nabla_y\left(g_{ec}(x,y)-\mathfrak{I}_{Q_t\times Q_s}[g_{ec}](x,y)\right)\|_2 \mid (x,y)\in Q_t\times Q_s\right\}\\&\leq\sqrt{3}\max_{i\in\underline{3}_{\mathrm{l}}}\left\{\left\|\tfrac{\partial}{\partial y_i}\left(g_{ec}-\mathfrak{I}_{Q_t\times Q_s}[g_{ec}]\right)\right\|_{\infty,Q_t\times Q_s}\right\}.\end{aligned}$$

Insgesamt gilt es somit eine Abschätzung für

$$\sqrt{3}\max_{i\in\underline{3}_{\mathrm{l}}}\left\{\left\|\tfrac{\partial}{\partial y_i}\left(g_{ec}-\mathfrak{I}_{Q_t\times Q_s}[g_{ec}]\right)\right\|_{\infty,Q_t\times Q_s}\right\}+\kappa\|g_{ec}-\mathfrak{I}_{Q_t\times Q_s}[g_{ec}]\|_{\infty,Q_t\times Q_s}$$

zu finden. Die zweite Norm ist nichts weiter als ein skalares Vielfaches des bereits untersuchten Fehlers des Einfachschichtoperators. Die Wellenzahl kann mit Hilfe der Zulässigkeitsbedingungen beschränkt werden, so dass sich die weitere Analyse auf die erste Norm konzentriert.
Die zu untersuchende Kernfunktion ergibt sich durch Ableiten der modifizierten Kernfunktion des Einfachschichtoperators und ist für $j\in\underline{3}_{\mathrm{l}}$ mit $j':=j+3$ durch

$$g_{dc,j'}(x,y)=g_{ec}(x,y)\left(\tfrac{\langle x-y,e_j\rangle_2}{\|x-y\|_2}\left(\tfrac{1}{\|x-y\|_2}-\mathrm{i}\kappa\right)+\mathrm{i}\kappa\langle c,e_j\rangle_2\right)\tag{4.4.1}$$

gegeben, deren Gegenstück mit eindimensionalem Definitionsbereich (4.2.4) für die Untersuchung des Interpolationsfehlers genutzt werden kann. Die Holomorphie dieser Funktion wurde schon im ersten Abschnitt des Kapitels zur Fehlerabschätzung gefolgert, so dass direkt mit der Analyse begonnen werden kann.
Wie beim Einfachschichtoperator soll der gesamte Interpolationsfehler auf die eindimensionalen Interpolationsfehler zurückgeführt werden, entsprechend soll der Satz 2.8 zu partiellen Ableitungen des Interpolationsoperators Anwendung finden. Die zu untersuchende Dimension ist durch $d=6$ gegeben, in Betracht kommen nur die folgenden drei Multiindizes $\alpha^{(1)}=(0,0,0,1,0,0)$, $\alpha^{(2)}=(0,0,0,0,1,0)$ und $\alpha^{(3)}=(0,0,0,0,0,1)$. Für ein $j\in\underline{3}_{\mathrm{l}}$, j' wie zuvor und $i\in\underline{6}_{\mathrm{l}}$ werden Abschätzungen

$$\left\|{}^{i}g_{dc,j'}-\mathfrak{I}_{[-1,1]}[{}^{i}g_{dc,j'}]\right\|_{\infty,[-1,1]}\leq\epsilon\qquad\text{falls } i\neq j'$$

beziehungsweise

$$\left\|\left({}^{i}g_{ec}-\mathfrak{I}_{[-1,1]}[{}^{i}g_{ec}]\right)'\right\|_{\infty,[-1,1]}\leq\widehat{\epsilon}\qquad\text{falls } i=j'$$

benötigt, um den Satz 2.8 anwenden zu können. Die Schranke für $i\neq j'$ kann leicht mit dem Satz 4.8 gewonnen werden. Das folgende Lemma leitet die für den Satz 4.8 notwendige Abschätzung von ${}^{i}g_{dc,j'}$ auf einer Bernstein-Ellipse her.

Lemma 4.40
Seien ein richtungsabhängiger Blockbaum $\mathcal{T}_{\mathcal{I}\times\mathcal{I}}$ sowie ein zulässiges Blatt $b = (t, s, c) \in \mathcal{T}_{\mathcal{I}\times\mathcal{I}}$, ein $i \in \underline{6}_1$ sowie $j' \in \underline{6}_{1\,4}$ mit $i \neq j'$ gegeben. Weiter seien $\mathfrak{m}, \mathfrak{l}$ und $\mathfrak{d}$ passend zur Interpolation gegeben. Setze für $r \in (0, \mathfrak{d})$ den Halbachsenparameter mit $\varrho = \sqrt{r^2+1}+r$. Dann kann die Kernfunktion des Doppelschichtoperators beschränkt werden durch

$$\max\left\{\left|{}^{i}g_{dc,j'}(z)\right| \mid z \in \overline{\mathcal{D}}_\varrho\right\} \leq \frac{\mathcal{C}_e^\star(\eta_1,\eta_2,r)}{\max\left\{\operatorname{diam}^2(Q_t),\operatorname{diam}^2(Q_s)\right\}}\left(\frac{\eta_2^2}{1-\frac{r}{\mathfrak{d}}}+2\eta_2\right),$$

wobei $\mathcal{C}_e^\star(\eta_1,\eta_2,r)$ wie in Lemma 4.18 definiert ist.

Beweis: Um eine obere Schranke zu finden, zerlege die Funktion ${}^{i}g_{dc,j'}$ und analysiere die einzelnen Teile. Eine Abschätzung für die Kernfunktion des Einfachschichtoperators ist schon aus Lemma 4.18 bekannt

$$\max\left\{|{}^{i}g_{ec}(z)| \mid z \in \overline{\mathcal{D}}_\varrho\right\} \leq \frac{\mathcal{C}_e^\star(\eta_1,\eta_2,r)}{\operatorname{dist}(Q_t,Q_s)}.$$

Für den übrig gebliebenen Faktor gilt für ein $z \in \overline{\mathcal{D}}_\varrho$

$$\left|\frac{\langle\mathfrak{m}-z\mathfrak{l},e_j\rangle_E}{\|\mathfrak{m}-z\mathfrak{l}\|_E}\left(\frac{1}{\|\mathfrak{m}-z\mathfrak{l}\|_E}-\mathrm{i}\kappa\right)+\mathrm{i}\kappa\langle c,e_j\rangle_2\right| \overset{\Delta}{\leq} \left|\frac{\langle\mathfrak{m}-z\mathfrak{l},e_j\rangle_E}{\|\mathfrak{m}-z\mathfrak{l}\|_E}\right|\left|\frac{1}{\|\mathfrak{m}-z\mathfrak{l}\|_E}-\mathrm{i}\kappa\right|+\left|\mathrm{i}\kappa\langle c,e_j\rangle_2\right|.$$

Sowohl $\frac{\langle\mathfrak{m}-z\mathfrak{l},e_j\rangle_E}{\|\mathfrak{m}-z\mathfrak{l}\|_E}$ als auch $\langle c,e_j\rangle_2$ haben einen Betrag, der kleiner gleich eins ist, entsprechend folgt

$$\left|\frac{\langle\mathfrak{m}-z\mathfrak{l},e_j\rangle_E}{\|\mathfrak{m}-z\mathfrak{l}\|_E}\right|\left|\frac{1}{\|\mathfrak{m}-z\mathfrak{l}\|_E}-\mathrm{i}\kappa\right|+\left|\mathrm{i}\kappa\langle c,e_j\rangle_2\right| \leq \left|\frac{1}{\|\mathfrak{m}-z\mathfrak{l}\|_E}-\mathrm{i}\kappa\right|+|\kappa| \overset{\Delta}{\leq} \left|\frac{1}{\|\mathfrak{m}-z\mathfrak{l}\|_E}\right|+2\kappa.$$

Weiterhin gilt mit Lemma 4.13 und demselben Vorgehen wie in Lemma 4.18

$$\left|\frac{1}{\|\mathfrak{m}-z\mathfrak{l}\|_E}\right| \leq \frac{1}{\|\mathfrak{l}\|_2(\mathfrak{d}-r)} \leq \frac{1}{\operatorname{dist}(Q_t,Q_s)(1-\frac{r}{\mathfrak{d}})}.$$

Da die Interpolation nur auf zulässigen Blöcken stattfindet, kann die Wellenzahl mit der parabolischen Zulässigkeitsbedingung (2.3.2b) weiter abgeschätzt werden

$$2\kappa \leq \frac{2\eta_2\operatorname{dist}(Q_t,Q_s)}{\max\left\{\operatorname{diam}^2(Q_t),\operatorname{diam}^2(Q_s)\right\}}.$$

Die Abschätzungen gelten ebenfalls für das Maximum, so dass

$$\begin{aligned}\max&\left\{\left|\frac{\langle\mathfrak{m}-z\mathfrak{l},e_j\rangle_E}{\|\mathfrak{m}-z\mathfrak{l}\|_E}\right|\left|\frac{1}{\|\mathfrak{m}-z\mathfrak{l}\|_E}-\mathrm{i}\kappa\right|+\left|\mathrm{i}\kappa\langle c,e_j\rangle_2\right| \mid z\in\overline{\mathcal{D}}_\varrho\right\}\\ &\leq \frac{1}{\operatorname{dist}(Q_t,Q_s)(1-\frac{r}{\mathfrak{d}})}+\frac{2\eta_2\operatorname{dist}(Q_t,Q_s)}{\max\left\{\operatorname{diam}^2(Q_t),\operatorname{diam}^2(Q_s)\right\}}\end{aligned}$$

folgt. Dies kann durch Umformen und Nutzen der Standardzulässigkeitsbedingung (2.3.2c) noch vereinfacht werden

$$\begin{aligned}
&\frac{1}{\operatorname{dist}(Q_t,Q_s)(1-\frac{r}{\delta})} + \frac{2\eta_2 \operatorname{dist}(Q_t,Q_s)}{\max\{\operatorname{diam}^2(Q_t),\operatorname{diam}^2(Q_s)\}} \\
= &\frac{1}{\max\{\operatorname{diam}^2(Q_t),\operatorname{diam}^2(Q_s)\}} \left(\frac{\max\{\operatorname{diam}^2(Q_t),\operatorname{diam}^2(Q_s)\}}{\operatorname{dist}(Q_t,Q_s)(1-\frac{r}{\delta})} + 2\eta_2 \operatorname{dist}(Q_t,Q_s) \right) \\
\overset{(2.3.2c)}{\leq} &\frac{1}{\max\{\operatorname{diam}^2(Q_t),\operatorname{diam}^2(Q_s)\}} \left(\frac{\eta_2^2 \operatorname{dist}^2(Q_t,Q_s)}{\operatorname{dist}(Q_t,Q_s)(1-\frac{r}{\delta})} + 2\eta_2 \operatorname{dist}(Q_t,Q_s) \right) \\
= &\frac{\operatorname{dist}(Q_t,Q_s)}{\max\{\operatorname{diam}^2(Q_t),\operatorname{diam}^2(Q_s)\}} \left(\frac{\eta_2^2}{1-\frac{r}{\delta}} + 2\eta_2 \right).
\end{aligned}$$

Die Multiplikation der beiden Teilabschätzungen führt zur Behauptung

$$\begin{aligned}
&\frac{\mathcal{C}_e^\star(\eta_1,\eta_2,r)}{\operatorname{dist}(Q_t,Q_s)} \frac{\operatorname{dist}(Q_t,Q_s)}{\max\{\operatorname{diam}^2(Q_t),\operatorname{diam}^2(Q_s)\}} \left(\frac{\eta_2^2}{1-\frac{r}{\delta}} + 2\eta_2 \right) \\
= &\frac{\mathcal{C}_e^\star(\eta_1,\eta_2,r)}{\max\{\operatorname{diam}^2(Q_t),\operatorname{diam}^2(Q_s)\}} \left(\frac{\eta_2^2}{1-\frac{r}{\delta}} + 2\eta_2 \right).
\end{aligned}$$

□

Für den Fall, dass $i = j'$ gilt, ist noch ein wenig mehr Vorarbeit nötig. Die Ableitung des Interpolationsfehlers soll mit Hilfe der Cauchy-Integralformel auf den Interpolationsfehler an sich zurückgeführt werden.
Für die Cauchy-Integralformel werden für alle $\tau \in [-1,1]$ Kreise $\{z \in \mathbb{C} \mid |z-\tau| = r\}$ betrachtet. Es stellt sich die Frage, wie der Radius zu wählen ist, wenn z vom Rand der Bernstein-Ellipse D_ϱ stammt. Da der Radius bei der Cauchy-Integralformel im Nenner auftaucht, wird der minimale Abstand vom Intervall $[-1,1]$ und dem Rand der Bernstein-Ellipse gesucht. Der Beweis stammt ursprünglich aus einer unveröffentlichten Arbeit von Herrn Börm [9].

Lemma 4.41 (Minimaler Radius)
Sei ein $\varrho \in \mathbb{R}_{>1}$ gegeben. Setze $r_{min} = a_\varrho - 1$, wobei a_ϱ die reelle Halbachse der Bernstein-Ellipse D_ϱ sei. Dann gilt

$$|z-\tau| \geq r_{min} \qquad \text{für alle } z \in \partial D_\varrho, \tau \in [-1,1].$$

Beweis: Vorweg mache eine Beobachtung, grundsätzlich gilt für alle $\tau \in [-1,1]$

$$|\tau+1| + |\tau-1| = \tau + 1 - \tau + 1 = 2.$$

Zeige die Behauptung per Kontraposition. Dazu seien $z \in \mathbb{C}$ und $\tau \in [-1,1]$ mit $|z-\tau| < r_{min}$ gegeben. Betrachte die Charakterisierung der Bernstein-Ellipse über die Brennpunkte,

um zu zeigen, dass $z \notin \partial D_\varrho$. Es gilt

$$\begin{aligned}|z+1|+|z-1| &= |z-\tau+\tau+1|+|z-\tau+\tau-1| \\ &\overset{\Delta}{\leq} |z-\tau|+(|\tau+1|+|\tau-1|)+|z-\tau| \\ &= |z-\tau|+2+|z-\tau| \\ &< r_{min}+2+r_{min} \\ &= 2+2a_\varrho-2 = 2a_\varrho.\end{aligned}$$

Damit folgt

$$|z+1|+|z-1| < 2a_\varrho,$$

was die Kontraposition zeigt. □

Lemma 4.42 (Ableitung des Interpolationsfehlers)
Seien ein richtungsabhängiger Blockbaum $\mathcal{T}_{\mathcal{I}\times\mathcal{I}}$ sowie ein zulässiges Blatt $b = (t,s,c) \in \mathcal{T}_{\mathcal{I}\times\mathcal{I}}$ gegeben. Sei $i \in \underline{6}_1$ gegeben und die Parameter $\mathfrak{m}, \mathfrak{l}$ und $\mathfrak{d}$ stammen aus der Interpolation von Ordnung $m \in \mathbb{N}_0$ auf diesem Block. Setze für $r \in (0,\mathfrak{d})$ den Parameter der Bernstein-Ellipse mit $\varrho = \sqrt{r^2+1}+r$ und wähle ein $\widehat{\varrho} \in (1,\varrho)$. Dann gilt

$$\left\| \left({}^i g_{ec} - \mathfrak{I}_{[-1,1]}[{}^i g_{ec}] \right)' \right\|_{\infty,[-1,1]} \leq \tfrac{1}{a_{\widehat{\varrho}}-1} \left(1+\Lambda_m\right) \tfrac{2}{\varrho\widehat{\varrho}^{-1}-1} \left(\tfrac{\widehat{\varrho}}{\varrho}\right)^m \left\| {}^i g_{ec} \right\|_{\infty,D_\varrho}.$$

Beweis: Setze $r_{min} = a_{\widehat{\varrho}} - 1$ und

$$h(\tau) = {}^i g_{ec}(\tau) - \mathfrak{I}_{[-1,1]}[{}^i g_{ec}](\tau) \qquad \text{für alle } \tau \in [-1,1].$$

Dann ist h als Differenz holomorpher Funktionen auch auf der Bernstein-Ellipse $D_{\widehat{\varrho}}$ holomorph, so dass sich mit der Cauchy-Integralformel für alle $s \in (0, r_{min})$

$$h'(\tau) = \tfrac{1}{2\mathrm{i}\pi} \int_{|z-\tau|=s} \tfrac{h(z)}{(z-\tau)^2} \mathrm{d}z$$

und damit

$$\begin{aligned}|h'(\tau)| &= \left| \tfrac{1}{2\mathrm{i}\pi} \int_{|z-\tau|=s} \tfrac{h(z)}{(z-\tau)^2} \mathrm{d}z \right| \overset{\Delta}{\leq} \tfrac{1}{2|\mathrm{i}|\pi} \int_{|z-\tau|=s} \tfrac{|h(z)|}{|z-\tau|^2} \mathrm{d}z \\ &= \tfrac{1}{2\pi} \int_{|z-\tau|=s} \tfrac{|h(z)|}{s^2} \mathrm{d}z \leq \tfrac{1}{2\pi} \int_{|z-\tau|=s} \tfrac{\|h\|_{\infty,D_{\widehat{\varrho}}}}{s^2} \mathrm{d}z\end{aligned}$$

ergibt. Der Integrand ist konstant, was zu

$$\tfrac{1}{2\pi} \int_{|z-\tau|=s} \tfrac{\|h\|_{\infty,D_{\widehat{\varrho}}}}{s^2} \mathrm{d}z = \tfrac{1}{2\pi} \tfrac{\|h\|_{\infty,D_{\widehat{\varrho}}}}{s^2} (2\pi s) = \tfrac{\|h\|_{\infty,D_{\widehat{\varrho}}}}{s}$$

führt. Eine Grenzwertbetrachtung für $s \to r_{min}$ liefert dann

$$\lim_{s \to r_{min}} \frac{\|h\|_{\infty,D_{\widehat{\varrho}}}}{s} = \frac{\|h\|_{\infty,D_{\widehat{\varrho}}}}{r_{min}} = \frac{\|h\|_{\infty,D_{\widehat{\varrho}}}}{a_{\widehat{\varrho}}-1}.$$

Ohne die Ableitung gilt es, nun noch für

$$\|^i g_{ec} - \mathfrak{I}_{[-1,1]}[^i g_{ec}]\|_{\infty,D_{\widehat{\varrho}}}$$

eine Abschätzung zu finden. Der Interpolationsfehler auf der Bernstein-Ellipse kann mit Lemma 4.9 durch

$$\left\|^i g_{ec} - \mathfrak{I}_{[-1,1]}[^i g_{ec}]\right\|_{\infty,D_{\widehat{\varrho}}} \leq (1+\Lambda_m)\, \tfrac{2}{\varrho\widehat{\varrho}^{-1}-1} \left(\tfrac{\widehat{\varrho}}{\varrho}\right)^m \|^i g_{ec}\|_{\infty,D_\varrho}$$

beschränkt werden, womit sich durch Zusammenfügen der Teilabschätzungen die Behauptung ergibt. □

Nehme an, dass Cluster in jeder Koordinate eine Ausdehnung größer null haben, so dass eine Konstante $\mathcal{C}_{ad} \in \mathbb{R}_{>1}$ existiert mit

$$\tfrac{\operatorname{diam}(Q_s)}{\operatorname{diam}(Q_{s,j})} \leq \mathcal{C}_{ad} \qquad \text{für alle } s \in \mathcal{T}_{\mathcal{I}}, j \in \underline{3}_{\mathrm{l}}. \tag{4.4.2}$$

Damit ist es möglich, den Approximationsfehler für den Doppelschichtoperator auf ähnliche Weise wie beim Einfachschichtoperator zu beschränken.

Satz 4.43
Seien ein richtungsabhängiger Blockbaum $\mathcal{T}_{\mathcal{I}\times\mathcal{I}}$ sowie ein zulässiges Blatt $b = (t,s,c) \in \mathcal{T}_{\mathcal{I}\times\mathcal{I}}$ gegeben. Weiter seien $\mathfrak{m}, \mathfrak{l}$ und $\mathfrak{d}$ passend zur Interpolation gegeben und die überdeckenden Quader erfüllen (4.4.2)*. Setze für $r \in (0,\mathfrak{d})$ den Halbachsenparameter mit $\varrho = \sqrt{r^2+1}+r$ und wähle ein $\widehat{\varrho} \in (1,\varrho)$. Der Fehler der Interpolation von Ordnung $m \in \mathbb{N}_0$ ist dann für alle $j \in \underline{3}_{\mathrm{l}}$ beschränkt durch*

$$\left\| \tfrac{\partial}{\partial y_j} \left(g_{ec} - \mathfrak{I}_{Q_t\times Q_s}[g_{ec}]\right)\right\|_{\infty,Q_t\times Q_s} \leq (1+\Lambda_m) \tfrac{10\Lambda_m^5}{\varrho\widehat{\varrho}^{-1}-1} \left(\tfrac{\widehat{\varrho}}{\varrho}\right)^m \tfrac{\eta_2 \mathcal{C}_e^\star(\eta_1,\eta_2,r)\mathcal{C}_d^\star(\eta_2,\widehat{\varrho},r,\mathfrak{d})}{\max\left\{\operatorname{diam}^2(Q_t),\operatorname{diam}^2(Q_s)\right\}},$$

wobei $\mathcal{C}_e^\star(\eta_1,\eta_2,r)$ wie in Lemma 4.18 und $\mathcal{C}_d^\star(\eta_2,\widehat{\varrho},r,\mathfrak{d})$ durch

$$\mathcal{C}_d^\star(\eta_2,\widehat{\varrho},r,\mathfrak{d}) := \frac{\mathcal{C}_{mk}\mathcal{C}_{ad}}{a_{\widehat{\varrho}}-1} + \tfrac{\eta_2}{1-\frac{r}{\mathfrak{d}}} + 2$$

mit den Konstanten $\mathcal{C}_{mk}$ nach (3.1.10) *und $\mathcal{C}_{ad}$ nach* (4.4.2) *definiert sind.*

Beweis: Um den Fehler abzuschätzen, soll der Satz 2.8 verwendet werden, wofür wiederum Abschätzungen für die eindimensionalen Interpolationsfehler mit $j \in \underline{3}_{\mathrm{l}}$ und $j' = j+3$

$$\left\|^i g_{dc,j'} - \mathfrak{I}_{[-1,1]}[^i g_{dc,j'}]\right\|_{\infty,[-1,1]} \qquad \text{für alle } i \neq j'$$

und

$$\left\| \left({}^{i}g_{ec} - \mathfrak{I}_{[-1,1]}[{}^{i}g_{ec}] \right)' \right\|_{\infty,[-1,1]} \qquad \text{für } i = j'$$

für alle $i \in \underline{6}$ notwendig sind. Im ersten Fall folgt mit dem Lemma 4.9

$$\left\| {}^{i}g_{dc,j'} - \mathfrak{I}_{[-1,1]}[{}^{i}g_{dc,j'}] \right\|_{\infty,[-1,1]} \leq (1+\Lambda_m) \tfrac{2}{\varrho-1} \varrho^{-m} \max\left\{ \left| {}^{i}g_{dc,j'}(z) \right| \mid z \in \overline{\mathcal{D}}_\varrho \right\}.$$

Das Maximum kann mit Lemma 4.40 abgeschätzt werden, so dass sich

$$\begin{aligned} &\left\| {}^{i}g_{dc,j'} - \mathfrak{I}_{[-1,1]}[{}^{i}g_{dc,j'}] \right\|_{\infty,[-1,1]} \\ &\leq (1+\Lambda_m) \tfrac{2}{\varrho-1} \varrho^{-m} \tfrac{\mathcal{C}_e^\star(\eta_1,\eta_2,r)}{\max\left\{\operatorname{diam}^2(Q_t),\operatorname{diam}^2(Q_s)\right\}} \left(\tfrac{\eta_2^2}{1-\frac{r}{\mathfrak{d}}} + 2\eta_2 \right) \end{aligned}$$

ergibt.
Der zweiten Fall kann mit dem gewählten $\widehat{\varrho} \in (1,\varrho)$ und dem Lemma 4.42 auf

$$\begin{aligned} \left\| \left({}^{i}g_{ec} - \mathfrak{I}_{[-1,1]}[{}^{i}g_{ec}] \right)' \right\|_{\infty,[-1,1]} &\leq \tfrac{1}{a_{\widehat{\varrho}}-1} (1+\Lambda_m) \tfrac{2}{\varrho\widehat{\varrho}^{-1}-1} \left(\tfrac{\widehat{\varrho}}{\varrho} \right)^m \left\| {}^{i}g_{ec} \right\|_{\infty,D_\varrho} \\ &\leq \tfrac{1}{a_{\widehat{\varrho}}-1} (1+\Lambda_m) \tfrac{2}{\varrho\widehat{\varrho}^{-1}-1} \left(\tfrac{\widehat{\varrho}}{\varrho} \right)^m \tfrac{\mathcal{C}_e^\star(\eta_1,\eta_2,r)}{\operatorname{dist}(Q_t,Q_s)} \end{aligned}$$

zurückgeführt werden. Im Satz 2.8 tritt noch die Inverse der Ableitung der Transformation auf. Für die Kernfunktionen ist die Transformation Φ für feste $x \in Q_t$ und $y \in Q_s$ mit $\mathfrak{m}, \mathfrak{l}$ wie zu Beginn des Kapitels 4 durch $\tau \mapsto \mathfrak{m} - \tau\mathfrak{l}$ für $\tau \in [-1,1]$ gegeben. Die Ableitung der Transformation erfüllt $\Phi' = -\mathfrak{l}$. Für $i = j+3$ ergibt sich mit der Definition von $\mathfrak{l}$ dann $|-\mathfrak{l}|^{-1} = \frac{2}{|b_{s,j}-a_{s,j}|}$ und entsprechend

$$\left| (\Phi')^{-1} \right| = \tfrac{2}{|b_{s,j}-a_{s,j}|} = \tfrac{2}{\operatorname{diam}(Q_{s,j})}.$$

Mit dem Satz 2.8 ergibt sich dann ein Fehler von

$$\begin{aligned} &(1+\Lambda_m) \tfrac{10\Lambda_m^4}{\varrho-1} \varrho^{-m} \tfrac{\mathcal{C}_e^\star(\eta_1,\eta_2,r)}{\max\left\{\operatorname{diam}^2(Q_t),\operatorname{diam}^2(Q_s)\right\}} \left(\tfrac{\eta_2^2}{1-\frac{r}{\mathfrak{d}}} + 2\eta_2 \right) \\ &+ (1+\Lambda_m) \tfrac{4\Lambda_m^5}{\varrho\widehat{\varrho}^{-1}-1} \left(\tfrac{\widehat{\varrho}}{\varrho} \right)^m \tfrac{\mathcal{C}_e^\star(\eta_1,\eta_2,r)}{\operatorname{dist}(Q_t,Q_s)} \tfrac{1}{\operatorname{diam}(Q_{s,j})} \tfrac{1}{a_{\widehat{\varrho}}-1}. \end{aligned}$$

Es gilt $\varrho^{-m} \leq \widehat{\varrho}^m \varrho^{-m}$ sowie $\varrho > \varrho\widehat{\varrho}^{-1} > 1$, so dass sich die Summe der Fehler durch

$$\begin{aligned} &(1+\Lambda_m) \tfrac{10\Lambda_m^5}{\varrho\widehat{\varrho}^{-1}-1} \left(\tfrac{\widehat{\varrho}}{\varrho} \right)^m \tfrac{\mathcal{C}_e^\star(\eta_1,\eta_2,r)}{\max\left\{\operatorname{diam}^2(Q_t),\operatorname{diam}^2(Q_s)\right\}} \\ &\left(\tfrac{\max\left\{\operatorname{diam}^2(Q_t),\operatorname{diam}^2(Q_s)\right\}}{\operatorname{dist}(Q_t,Q_s)\operatorname{diam}(Q_{s,j})(a_{\widehat{\varrho}}-1)} + \left(\tfrac{\eta_2^2}{1-\frac{r}{\mathfrak{d}}} + 2\eta_2 \right) \right) \end{aligned}$$

beschränken lässt. Mit der Standardzulässigkeitsbedingung (2.3.2c) und der Konstante $\mathcal{C}_{mk}$ nach (3.1.10), die das Verhältnis der Durchmesser von zwei Clustern t, s auf einer Stufe zueinander beschreibt, also für $\ell = \operatorname{stufe}(t)$

$$\operatorname{diam}(Q_t) \leq \operatorname{diam}_{\max}(\ell) \leq \mathcal{C}_{mk} \operatorname{diam}_{\min}(\ell) \leq \mathcal{C}_{mk} \operatorname{diam}(Q_s)$$

liefert, kann zusammen mit (4.4.2) weiter abgeschätzt werden

$$\begin{aligned}\left(\frac{\max\{\operatorname{diam}^2(Q_t),\operatorname{diam}^2(Q_s)\}}{\operatorname{diam}(Q_{s,j})(a_{\widehat{\varrho}}-1)\operatorname{dist}(Q_t,Q_s)}+\frac{\eta_2^2}{1-\frac{r}{\mathfrak{d}}}+2\eta_2\right) &\leq \eta_2\left(\frac{\max\{\operatorname{diam}(Q_t),\operatorname{diam}(Q_s)\}}{\operatorname{diam}(Q_{s,j})(a_{\widehat{\varrho}}-1)}+\frac{\eta_2}{1-\frac{r}{\mathfrak{d}}}+2\right)\\ &\leq \eta_2\left(\frac{\operatorname{diam}(Q_s)}{\operatorname{diam}(Q_{s,j})}\frac{\mathcal{C}_{mk}}{a_{\widehat{\varrho}}-1}+\frac{\eta_2}{1-\frac{r}{\mathfrak{d}}}+2\right)\\ &\leq \eta_2\left(\mathcal{C}_{ad}\frac{\mathcal{C}_{mk}}{a_{\widehat{\varrho}}-1}+\frac{\eta_2}{1-\frac{r}{\mathfrak{d}}}+2\right),\end{aligned}$$

so dass mit $\mathcal{C}_d^\star(\eta_2,\widehat{\varrho},r,\mathfrak{d})$ die Behauptung folgt. □

Theorem 4.44 (Approximationsfehler)
Seien ein richtungsabhängiger Blockbaum $\mathcal{T}_{\mathcal{I}\times\mathcal{I}}$ sowie ein zulässiges Blatt $b=(t,s,c)\in\mathcal{T}_{\mathcal{I}\times\mathcal{I}}$ gegeben. Weiter seien $\mathfrak{d}$ passend zur Interpolation von Ordnung $m\in\mathbb{N}_0$ und ein $r\in(0,\mathfrak{d})$, $\varrho=\sqrt{r^2+1}+r$ und $\widehat{\varrho}\in(1,\varrho)$ gegeben und die überdeckenden Quader erfüllen (4.4.2)*. Dann ist der Interpolationsfehler des Doppelschichtoperators beschränkt durch*

$$\begin{aligned}\left\|\tfrac{\partial}{\partial n(y)}\left(g_e-\mathfrak{I}^c_{Q_t\times Q_s}[g_e]\right)\right\|_{\infty,Q_t\times Q_s} &\leq (1+\Lambda_m)\tfrac{10\sqrt{3}\Lambda_m^5}{\varrho\widehat{\varrho}^{-1}-1}\left(\tfrac{\widehat{\varrho}}{\varrho}\right)^m\tfrac{\eta_2\mathcal{C}_e^\star(\eta_1,\eta_2,r)}{\max\{\operatorname{diam}^2(Q_t),\operatorname{diam}^2(Q_s)\}}\\ &\quad\cdot\left(\mathcal{C}_d^\star(\eta_2,\widehat{\varrho},r,\mathfrak{d})+1\right),\end{aligned}$$

wobei $\mathcal{C}_e^\star(\eta_1,\eta_2,r)$ wie in Lemma 4.18 und $\mathcal{C}_d^\star(\eta_2,\widehat{\varrho},r,\mathfrak{d})$ wie in Satz 4.43 definiert sind.

Beweis: Nach der anfänglichen Betrachtung kann der Fehler zerlegt werden

$$\begin{aligned}\left\|\tfrac{\partial}{\partial n(y)}\left(g_e-\mathfrak{I}^c_{Q_t\times Q_s}[g_e]\right)\right\|_{\infty,Q_t\times Q_s} &\leq \sqrt{3}\max_{j\in\mathfrak{Z}}\left\{\left\|\tfrac{\partial}{\partial y_j}\left(g_{ec}-\mathfrak{I}_{Q_t\times Q_s}[g_{ec}]\right)\right\|_{\infty,Q_t\times Q_s}\right\}\\ &\quad+\kappa\|g_{ec}-\mathfrak{I}_{Q_t\times Q_s}[g_{ec}]\|_{\infty,Q_t\times Q_s}.\end{aligned}$$

Mit Satz 4.43 kann der erste Summand durch

$$\begin{aligned}&\sqrt{3}\max_{j\in\mathfrak{Z}}\left\{\left\|\tfrac{\partial}{\partial y_j}\left(g_{ec}-\mathfrak{I}_{Q_t\times Q_s}[g_{ec}]\right)\right\|_{\infty,Q_t\times Q_s}\right\}\leq(1+\Lambda_m)\tfrac{10\sqrt{3}\Lambda_m^5}{\varrho\widehat{\varrho}^{-1}-1}\left(\tfrac{\widehat{\varrho}}{\varrho}\right)^m\\ &\cdot\tfrac{\eta_2\mathcal{C}_e^\star(\eta_1,\eta_2,r)\mathcal{C}_d^\star(\eta_2,\widehat{\varrho},r,\mathfrak{d})}{\max\{\operatorname{diam}^2(Q_t),\operatorname{diam}^2(Q_s)\}}\end{aligned}$$

beschränkt werden, während der zweite Summand mit der Fehleraussage vom Einfachschichtoperator abgeschätzt werden kann. Dazu verwende die Schranke aus Theorem 4.19, bei der jedoch das $\varrho-1$ im Nenner noch nicht durch r abgeschätzt wurde. Dann ist der zweite Summand durch

$$\kappa\,\|g_{ec}-\mathfrak{I}_{Q_t\times Q_s}[g_{ec}]\|_{\infty,Q_t\times Q_s}\leq\kappa(1+\Lambda_m)\tfrac{12\Lambda_m^5}{\operatorname{dist}(Q_t,Q_s)}\varrho^{-m}\tfrac{\mathcal{C}_e^\star(\eta_1,\eta_2,r)}{\varrho-1}$$

beschränkt. Erneut kann die Wellenzahl mit der parabolischen Zulässigkeitsbedingung abgeschätzt werden, so dass

$$\kappa\leq\tfrac{\eta_2\operatorname{dist}(Q_t,Q_s)}{\max\{\operatorname{diam}^2(Q_t),\operatorname{diam}^2(Q_s)\}}$$

gilt und zusammen mit $12 < 10\sqrt{3}$ folgt dann

$$\begin{aligned}&\kappa(1+\Lambda_m)\tfrac{12\Lambda_m^5}{\operatorname{dist}(Q_t,Q_s)}\varrho^{-m}\tfrac{\mathcal{C}_e^\star(\eta_1,\eta_2,r)}{\varrho-1}\\&\leq(1+\Lambda_m)\tfrac{10\sqrt{3}\Lambda_m^5}{\varrho-1}\varrho^{-m}\tfrac{\eta_2\mathcal{C}_e^\star(\eta_1,\eta_2,r)}{\max\{\operatorname{diam}^2(Q_t),\operatorname{diam}^2(Q_s)\}}.\end{aligned}$$

Mit $\varrho^{-1}\leq\widehat{\varrho}\underline{\varrho}^{-1}$ sowie $\varrho>\varrho\widehat{\underline{\varrho}}^{-1}>1$ gilt

$$\begin{aligned}&(1+\Lambda_m)\tfrac{10\sqrt{3}\Lambda_m^5}{\varrho-1}\varrho^{-m}\tfrac{\eta_2\mathcal{C}_e^\star(\eta_1,\eta_2,r)}{\max\{\operatorname{diam}^2(Q_t),\operatorname{diam}^2(Q_s)\}}\\&\leq(1+\Lambda_m)\tfrac{10\sqrt{3}\Lambda_m^5}{\varrho\widehat{\underline{\varrho}}^{-1}-1}\left(\tfrac{\widehat{\underline{\varrho}}}{\varrho}\right)^m\tfrac{\eta_2\mathcal{C}_e^\star(\eta_1,\eta_2,r)}{\max\{\operatorname{diam}^2(Q_t),\operatorname{diam}^2(Q_s)\}}.\end{aligned}$$

Insgesamt ergibt sich damit

$$\begin{aligned}\left\|\tfrac{\partial}{\partial n(y)}\left(g_e-\mathfrak{I}^c_{Q_t\times Q_s}[g_e]\right)\right\|_{\infty,Q_t\times Q_s}&\leq(1+\Lambda_m)\tfrac{10\sqrt{3}\Lambda_m^5}{\varrho\widehat{\underline{\varrho}}^{-1}-1}\left(\tfrac{\widehat{\underline{\varrho}}}{\varrho}\right)^m\tfrac{\eta_2\mathcal{C}_e^\star(\eta_1,\eta_2,r)}{\max\{\operatorname{diam}^2(Q_t),\operatorname{diam}^2(Q_s)\}}\\&\quad\cdot\left(\mathcal{C}_d^\star(\eta_2,\widehat{\varrho},r,\mathfrak{d})+1\right).\end{aligned}$$

□

Auch beim Doppelschichtoperator kann die Fehlerabschätzung durch eine Zusatzannahme weiter vereinfacht werden.

Korollar 4.45
Seien die Voraussetzungen von Theorem 4.44 erfüllt und es gelte $r=\min\left\{1,\frac{3}{4}\mathfrak{d}\right\}$. *Dann kann eine knappere und damit handlichere Fehlerdarstellung*

$$\begin{aligned}&\left\|\tfrac{\partial}{\partial n(y)}\left(g_e-\mathfrak{I}^c_{Q_t\times Q_s}[g_e]\right)\right\|_{\infty,Q_t\times Q_s}\leq(1+\Lambda_m)\tfrac{10\sqrt{3}\Lambda_m^5}{\varrho\widehat{\underline{\varrho}}^{-1}-1}\left(\tfrac{\widehat{\underline{\varrho}}}{\varrho}\right)^m\\&\tfrac{\eta_2e^{\eta_1+\eta_2}}{\pi\max\{\operatorname{diam}^2(Q_t),\operatorname{diam}^2(Q_s)\}}\left(\tfrac{\mathcal{C}_{mk}\mathcal{C}_{ad}}{a_{\widehat{\underline{\varrho}}^{-1}}}+4\eta_2+3\right)\end{aligned}$$

erhalten werden.

Beweis: Die Einschränkung von r auf $\min\left\{1,\frac{3}{4}\mathfrak{d}\right\}$ ermöglicht es, das Korollar 4.20 aus dem Fall des Einfachschichtoperators zu nutzen und damit $\mathcal{C}_e^\star(\eta_1,\eta_2,r)$ zu beschränken. Außerdem kann auch $\mathcal{C}_d^\star(\eta_2,\widehat{\varrho},r,\mathfrak{d})$ mit $(1-\frac{r}{\mathfrak{d}})^{-1}\leq 4$ weiter vereinfacht werden, entsprechend folgt

$$\mathcal{C}_d^\star(\eta_2,\widehat{\varrho},r,\mathfrak{d})+1\leq\tfrac{\mathcal{C}_{mk}\mathcal{C}_{ad}}{a_{\widehat{\underline{\varrho}}^{-1}}}+4\eta_2+3.$$

□

Als praktisch erweist sich, dass die Aussagen für den Doppelschichtoperator leicht auf den adjungierten Doppelschichtoperator übertragen werden können. Gesucht wird in diesem

Fall eine Fehlerabschätzung für

$$\begin{aligned}&\left\| \tfrac{\partial}{\partial n(x)} \left(g_e - \mathfrak{I}^c_{Q_t \times Q_s}[g_e] \right) \right\|_{\infty, Q_t \times Q_s} \\ &\leq \left(\sqrt{3} \max_{i \in \underline{3}} \left\{ \left\| \tfrac{\partial}{\partial x_i} \left(g_{ec} - \mathfrak{I}_{Q_t \times Q_s}[g_{ec}] \right) \right\|_{\infty, Q_t \times Q_s} \right\} + \kappa \| g_{ec} - \mathfrak{I}_{Q_t \times Q_s}[g_{ec}] \|_{\infty, Q_t \times Q_s} \right).\end{aligned}$$

Die zu untersuchende Kernfunktion ergibt sich für $j \in \underline{3}$

$$g_{adc,j}(x,y) := \tfrac{\partial}{\partial x_j} g_{ec}(x,y) = g_{ec}(x,y) \left(\tfrac{\langle x-y, e_j \rangle_2}{\|x-y\|_2} \left(\mathrm{i}\kappa - \tfrac{1}{\|x-y\|_2} \right) - \mathrm{i}\kappa \langle c, e_j \rangle_2 \right),$$

wobei $e_j \in \mathbb{R}^3$ der j-te kanonische Einheitsvektor sei. In Betracht kommen die Multiindizes

$$\alpha \in \{(1,0,0,0,0,0), (0,1,0,0,0,0), (0,0,1,0,0,0)\}.$$

4.4.1 Reinterpolation

Wie schon zuvor beim Einfachschichtoperator muss die auftretende Reinterpolation und der damit einhergehende Fehler untersucht werden.
Auch in diesem Fall ermöglicht die Tensorinterpolation eine Rückführung des Fehlers auf Betrachtungen im eindimensionalen Fall. Die Notation der auftretenden Größen entspricht der im Kapitel zur Reinterpolation beim Einfachschichtoperator (siehe Kapitel 4.3.1).
Betrachtet werden durch Sequenzen $\mathfrak{s}_t^L, \mathfrak{s}_s^L$ dargestellte Pfade von Clustern sowie die dazugehörigen Sequenzen an Richtungen $\mathfrak{s}_c^L$. Auch bei der Reinterpolation kann die Norm des Fehlers durch Ausschreiben der Normalenableitung und Ausnutzen der Äquivalenz der euklidischen und Maximum-Norm (siehe Kapitel 4.4) umgeformt werden, dieses Mal jedoch ohne die ebene Welle zu eliminieren

$$\begin{aligned}&\left\| \tfrac{\partial}{\partial n(y)} \left(g_e - \mathfrak{I}^{c_L}_{Q_{t_L} \times Q_{s_L}} \circ \cdots \circ \mathfrak{I}^{c_0}_{Q_{t_0} \times Q_{s_0}}[g_e] \right) \right\|_{\infty, Q_{t_L} \times Q_{s_L}} \\ &\qquad \leq \sqrt{3} \max \left\{ \left\| \tfrac{\partial}{\partial y_j} \left(g_e - \mathfrak{I}^{c_L}_{Q_{t_L} \times Q_{s_L}} \circ \cdots \circ \mathfrak{I}^{c_0}_{Q_{t_0} \times Q_{s_0}}[g_e] \right) \right\|_{\infty, Q_{t_L} \times Q_{s_L}} \mid j \in \underline{3} \right\}.\end{aligned}$$

Da auch die partielle Ableitung einen linearen Operator bildet, kann wie in Kapitel 4.3.1 eine Teleskopsumme genutzt werden, um den Fehler in ein Produkt der Operatornorm und des Fehlers nach einmaliger Interpolation aufzuspalten. Entsprechend reicht es, für $\ell \in \underline{L}$ Interpolationsoperatoren

$$\tfrac{\partial}{\partial y_j} \circ \mathfrak{I}^{c_\ell}_{Q_{t_\ell} \times Q_{s_\ell}} = \tfrac{\partial}{\partial y_j} \circ \left(\mathfrak{I}^{c_\ell}_{Q_{t_\ell}} \otimes \mathfrak{I}^{-c_\ell}_{Q_{s_\ell}} \right)$$

für $j \in \underline{3}$ zu untersuchen. Die Interpolationsoperatoren können so umsortiert werden, dass alle zu Clustern aus $\mathfrak{s}_t^L$ beziehungsweise $\mathfrak{s}_s^L$ gehörenden Operatoren zusammen stehen. Da

die Ableitungen nur den Interpolationsoperator für die Cluster in $\mathfrak{s}_s^L$ betreffen, kann die Ableitung im Tensorprodukt an den Clustern $\mathfrak{s}_t^L$ vorbeigezogen werden

$$\begin{aligned}\tfrac{\partial}{\partial y_j}\circ\left(\mathfrak{I}_{Q_{t_L}}^{c_L}\otimes\mathfrak{I}_{Q_{s_L}}^{-c_L}\circ\cdots\circ\mathfrak{I}_{Q_{t_1}}^{c_1}\otimes\mathfrak{I}_{Q_{s_1}}^{-c_1}\right)\\ =\left(\mathfrak{I}_{Q_{t_L}}^{c_L}\circ\cdots\circ\mathfrak{I}_{Q_{t_1}}^{c_1}\right)\otimes\left(\tfrac{\partial}{\partial y_j}\circ\left(\mathfrak{I}_{Q_{s_L}}^{-c_L}\circ\cdots\circ\mathfrak{I}_{Q_{s_1}}^{-c_1}\right)\right).\end{aligned}$$

Für die Fehlerbetrachtung kann der Interpolationsoperator für die Cluster in $\mathfrak{s}_t^L$ zunächst ignoriert werden, da er schon im Fall des Einfachschichtoperators untersucht wurde. Selbes gilt für die zwei Dimensionen von y, in deren Richtung nicht abgeleitet wird, so dass sich die Untersuchung als Erstes für ein festes $j\in\underline{3}$ und $i:=j+3$ auf

$$\tfrac{d}{dy_j}\circ\left(\mathfrak{I}_{I_L}^{i,c_L}\circ\cdots\circ\mathfrak{I}_{I_1}^{i,c_1}\right)$$

konzentriert. Die zu untersuchenden Intervalle werden erneut durch $I_{\ell,i}$ für $\ell\in\underline{L}$ und $i\in\underline{6}$ beziehungsweise durch I_ℓ bezeichnet (siehe (4.3.4)). Das erste Ziel ist, eine Aussage über den Fehler

$$\left\|\tfrac{d}{dy_j}\left(e^{\mathrm{i}\kappa\dot{c}_{\ell-1}^i\cdot}p-\mathfrak{I}_{I_\ell}^{i,c_\ell}[e^{\mathrm{i}\kappa\dot{c}_{\ell-1}^i\cdot}p]\right)\right\|_{\infty,I_\ell},$$

wobei $p\in\Pi_m$ ein Polynom sei, zu erhalten. Für die theoretische Betrachtung wird eine allgemeinere Form mit einer holomorphen Funktion f bewiesen.
Um die Ableitung zu eliminieren, soll erneut die Cauchy-Integralformel verwendet werden. Beim Einsatz der Cauchy-Integralformel ist es nötig, den Radius der betrachteten Kreisscheiben abzuschätzen. Da die Kreisscheiben innerhalb einer Bernstein-Ellipse liegen, hängt der minimale Radius von der umfassenden Bernstein-Ellipse ab (siehe Lemma 4.41). Wird eine auf ein Intervall $I_{\ell,i}$ transformierte Bernstein-Ellipse ${}^{I_{\ell,i}}D_\varrho$ genutzt, muss auch der minimale Radius mit skaliert werden. Mit der Abbildung $\Phi_{I_{\ell,i}}$ (siehe (4.3.6)) lässt sich der skalierte Radius leicht bestimmen. Dazu sei $m_{I_{\ell,i}}$ der Mittelpunkt des Intervalls $I_{\ell,i}$. Nach Lemma 4.41 gilt für alle $\widehat{z}\in\partial D_\varrho$ und $\widehat{\tau}\in[-1,1]$, dass $|\widehat{z}-\widehat{\tau}|\geq a_\varrho-1=r_{min}$ erfüllt. Werden nun ein $z\in\partial({}^{I_{\ell,i}}D_\varrho)$ und ein $\tau\in I_{\ell,i}$ betrachtet, existieren Urbilder $\widehat{z}\in\partial D_\varrho$ und $\widehat{\tau}\in[-1,1]$ mit

$$|z-\tau|=\left|\Phi_{I_{\ell,i}}(\widehat{z})-\Phi_{I_{\ell,i}}(\widehat{\tau})\right|=\left|m_{I_{\ell,i}}+\tfrac{|I_{\ell,i}|}{2}\widehat{z}-\left(m_{I_{\ell,i}}+\tfrac{|I_{\ell,i}|}{2}\widehat{\tau}\right)\right|=\tfrac{|I_{\ell,i}|}{2}|\widehat{z}-\widehat{\tau}|,$$

so dass sich bei der transformierten Bernstein-Ellipse

$$|z-\tau|\geq\tfrac{|I_{\ell,i}|}{2}(a_\varrho-1)\tag{4.4.3}$$

ergibt.

Lemma 4.46 (Differenzierte einmalige Reinterpolation)
Seien für $L\in\underline{p_{\mathcal{I}}}$ und ein $i\in\underline{6}$ eine Intervallsequenz, welche die Kontraktionseigenschaft

(4.3.5) *erfüllt, ein* $\varrho \in \mathbb{R}_{>1}$ *sowie* $\sigma := \vartheta(\varrho)$ *gegeben und es existiere ein* $\gamma_\varrho \in \mathbb{R}_{>0}$*, das* (4.3.10) *erfüllt. Weiter sei der betrachtete Interpolationsoperator von Ordnung* $m \in \mathbb{N}_0$ *stabil nach Bemerkung 6. Dann gibt es zu jedem* $q \in (\sigma^{-1}, 1)$ *ein* $c_{q,\varrho}$ *wie in Lemma 4.24, mit dem für alle* $\ell \in \underline{L}_\sqcup$ *mit* $I_\ell = I_{\ell,i}$ *und für alle auf* ${}^{I_{\ell-1}}D_\varrho$ *holomorphen Funktionen* f

$$\left\| \left(f - \mathfrak{I}_{I_\ell}^{i,c_\ell}[f] \right)' \right\|_{\infty, I_\ell} \leq \tfrac{2}{|I_\ell|} \tfrac{\omega_{I_{\ell,i},\varrho}}{a_\varrho - 1} e^{\gamma_\varrho} c_{q,\varrho} q^m \| e^{-\mathrm{i}\kappa \dot{c}_0^i \cdot} f \|_{\infty, {}^{I_{\ell-1}}D_\varrho}$$

gilt, wobei $\omega_{I_{\ell,i},\varrho}$ *für alle* $i \in \underline{6}_\sqcup$ *und* $\ell \in \underline{L}_\sqcup$ *durch*

$$\omega_{I_{\ell,i},\varrho} := \| e^{\mathrm{i}\kappa \dot{c}_\ell^i \cdot} \|_{\infty, {}^{I_{\ell,i}}D_\varrho}$$

gegeben sei.

Beweis: Seien ein $\ell \in \underline{L}_\sqcup$ und eine auf ${}^{I_{\ell-1}}D_\varrho$ holomorphe Funktionen f gegeben. Setze $h := f - \mathfrak{I}_{I_\ell}^{i,c_\ell}[f]$, dann ist h als Summe einer holomorphen Funktion f und dem Produkt von einem Polynom und einer ebenen Wellen auch holomorph auf ${}^{I_\ell}D_\varrho$. Verwende die Cauchy-Integralformel und erhalte

$$h'(\tau) = \tfrac{1}{2\mathrm{i}\pi} \int_{|z-\tau|=r} \tfrac{h(z)}{(z-\tau)^2} \, \mathrm{d}z \qquad \text{für alle } \tau \in I_\ell.$$

Entsprechend folgt für alle $\tau \in I_\ell$ und $r \in (0, r_{min})$ zusammen mit der transformierten Bernstein-Ellipse ${}^{I_\ell}D_\varrho$

$$\begin{aligned} |h'(\tau)| &= \left| \tfrac{1}{2\mathrm{i}\pi} \int_{|z-\tau|=r} \tfrac{h(z)}{(z-\tau)^2} \, \mathrm{d}z \right| \overset{\Delta}{\leq} \tfrac{1}{2\pi} \int_{|z-\tau|=r} \left| \tfrac{h(z)}{(z-\tau)^2} \right| \, \mathrm{d}z \\ &\leq \tfrac{1}{2\pi} \int_{|z-\tau|=r} \frac{\|h\|_{\infty, {}^{I_\ell}D_\varrho}}{r^2} \, \mathrm{d}z \\ &= \tfrac{1}{2\pi} (2\pi r) \frac{\|h\|_{\infty, {}^{I_\ell}D_\varrho}}{r^2}. \end{aligned}$$

Mit dem transformierten minimalen Radius $r_{min} = 2^{-1}|I_\ell|(a_\varrho - 1)$ nach (4.4.3) ergibt sich

$$\lim_{r \to r_{min}} \frac{\|h\|_{\infty, {}^{I_\ell}D_\varrho}}{r} = \frac{\|h\|_{\infty, {}^{I_\ell}D_\varrho}}{2^{-1}|I_\ell|(a_\varrho - 1)}.$$

Nachdem die Ableitung eliminiert wurde, bleibt ein bekannter Fehler übrig. Da der Fehler nun aber auf einer Bernstein-Ellipse betrachtet wird, kann das Lemma 4.26 nicht direkt angewendet werden. Jedoch lässt sich der Ausdruck so umformen, dass auf das Lemma 4.24 zurückgegriffen werden kann. Dazu ziehe zunächst eine ebene Welle aus der Norm heraus

$$\begin{aligned} \| f - \mathfrak{I}_{I_\ell}^{i,c_\ell}[f] \|_{\infty, {}^{I_\ell}D_\varrho} &= \| e^{\mathrm{i}\kappa \dot{c}_\ell^i \cdot} e^{-\mathrm{i}\kappa \dot{c}_\ell^i \cdot} \left(f - \mathfrak{I}_{I_\ell}^{i,c_\ell}[f] \right) \|_{\infty, {}^{I_\ell}D_\varrho} \\ &\leq \| e^{\mathrm{i}\kappa \dot{c}_\ell^i \cdot} \|_{\infty, {}^{I_\ell}D_\varrho} \| e^{-\mathrm{i}\kappa \dot{c}_\ell^i \cdot} f - \mathfrak{I}_{I_\ell}[e^{-\mathrm{i}\kappa \dot{c}_\ell^i \cdot} f] \|_{\infty, {}^{I_\ell}D_\varrho}. \end{aligned}$$

Auf diesen Fehler kann das Lemma 4.24 angewendet werden

$$\|e^{-\mathrm{i}\kappa\dot{c}_\ell^i\cdot} f - \mathfrak{I}_{I_\ell}[e^{-\mathrm{i}\kappa\dot{c}_\ell^i\cdot} f]\|_{\infty,{}^{I_\ell}D_\varrho} \leq c_{q,\varrho} q^m \|e^{-\mathrm{i}\kappa\dot{c}_\ell^i\cdot} f\|_{\infty,{}^{I_{\ell-1}}D_\varrho}.$$

Anschließend wird noch die ebene Welle entlang der Richtung $\dot{c}_0^i$ eingeschoben und mit (4.3.11) abgeschätzt

$$\|e^{-\mathrm{i}\kappa\dot{c}_\ell^i\cdot} e^{\mathrm{i}\kappa\dot{c}_0^i\cdot} e^{-\mathrm{i}\kappa\dot{c}_0^i\cdot} f\|_{\infty,{}^{I_{\ell-1}}D_\varrho} \leq e^{\gamma_\varrho} \|e^{-\mathrm{i}\kappa\dot{c}_0^i\cdot} f\|_{\infty,{}^{I_{\ell-1}}D_\varrho}.$$

Durch Zusammenfügen der Teilergebnisse folgt die Behauptung

$$\left\| \left(f - \mathfrak{I}_{I_\ell}^{i,c_\ell}[f] \right)' \right\|_{\infty,I_\ell} \leq \tfrac{2}{|I_\ell|} \tfrac{\omega_{I_{\ell,i},\varrho}}{a_\varrho - 1} e^{\gamma_\varrho} c_{q,\varrho} q^m \|e^{-\mathrm{i}\kappa\dot{c}_0^i\cdot} f\|_{\infty,{}^{I_{\ell-1}}D_\varrho}.$$

□

Die Richtung in der ebenen Welle auf der rechten Seite der Fehleraussagen in Lemma 4.46 kann durch eine beliebige Richtung $\dot{c}_k^i$ für $k \in \underline{\ell}_\rfloor$ ersetzt werden.

Um die Lesbarkeit der Beweise deutlich zu erleichtern, betrachte zwischenzeitlich eine gewichtete Maximum-Norm. Definiere die mit einer ebenen Welle gewichtete Maximum-Norm auf D_ϱ für alle Funktionen $u \in C(D_\varrho)$ über

$$\|u\|_{\widehat{\infty}_{\ell,i},D_\varrho} := \|e^{-\mathrm{i}\kappa\dot{c}_\ell^i\cdot} u\|_{\infty,D_\varrho} \qquad \text{für alle } \ell \in \underline{L}_{\rfloor 0}, i \in \underline{6}_\rfloor.$$

Da die ebene Welle auf einer Bernstein-Ellipse nicht null wird, übertragen sich die Normeigenschaften der Maximum-Norm auf die gewichtete Maximum-Norm.
Mit Hilfe der gewichteten Maximum-Norm beweise zunächst eine für den Fehler der Reinterpolation wichtige Hilfsaussage.

Lemma 4.47 (Normabschätzung)
Seien für $L \in \underline{p_{\mathcal{I}}}_\rfloor$ und ein $i \in \underline{6}_\rfloor$ eine Intervallsequenz, welche die Kontraktionseigenschaft (4.3.5) *erfüllt, ein $\varrho \in \mathbb{R}_{>1}$ sowie $\sigma := \vartheta(\varrho)$ gegeben und es existiere ein $\gamma_\varrho \in \mathbb{R}_{>0}$, das* (4.3.10) *erfüllt. Für alle $\ell \in \underline{L}_\rfloor$ mit $I_\ell = I_{\ell,i}$ und eine auf ${}^{I_0}D_\varrho$ holomorphe Funktion f gilt*

$$\|f\|_{\widehat{\infty}_{\ell,i},{}^{I_\ell}D_\varrho} \leq e^{\gamma_\varrho} \|f\|_{\widehat{\infty}_{\ell-n,i},{}^{I_{\ell-n}}D_\varrho} \qquad \textit{für alle } n \in \underline{\ell}_\rfloor.$$

Beweis: Seien ein $\ell \in \underline{L}_\rfloor$ und $n \in \underline{\ell}_\rfloor$ gegeben. Verwende die Bedingung (4.3.10), so dass

$$\begin{aligned}
\|f\|_{\widehat{\infty}_{\ell,i},{}^{I_\ell}D_\varrho} = \|e^{-\mathrm{i}\kappa\dot{c}_\ell^i\cdot} f\|_{\infty,{}^{I_\ell}D_\varrho} &= \|e^{-\mathrm{i}\kappa\dot{c}_\ell^i\cdot} e^{\mathrm{i}\kappa\dot{c}_{\ell-n}^i\cdot} e^{-\mathrm{i}\kappa\dot{c}_{\ell-n}^i\cdot} f\|_{\infty,{}^{I_\ell}D_\varrho} \\
&\leq \|e^{\mathrm{i}\kappa(\dot{c}_{\ell-n}^i - \dot{c}_\ell^i)\cdot}\|_{\infty,{}^{I_\ell}D_\varrho} \|e^{-\mathrm{i}\kappa\dot{c}_{\ell-n}^i\cdot} f\|_{\infty,{}^{I_\ell}D_\varrho} \\
&\leq e^{\gamma_\varrho} \|e^{-\mathrm{i}\kappa\dot{c}_{\ell-n}^i\cdot} f\|_{\infty,{}^{I_\ell}D_\varrho}
\end{aligned}$$

folgt. Mit dem Lemma 4.23 zu den geschachtelten Bernstein-Ellipsen folgt dann die Behauptung

$$\begin{aligned} e^{\gamma_\varrho}\|e^{-\mathrm{i}\kappa\dot{c}^i_{\ell-n}\cdot} f\|_{\infty,\,{}^{I_\ell}D_\varrho} &\le e^{\gamma_\varrho}\|e^{-\mathrm{i}\kappa\dot{c}^i_{\ell-n}\cdot} f\|_{\infty,\,{}^{I_\ell}D_{\sigma^n\varrho}} \le e^{\gamma_\varrho}\|e^{-\mathrm{i}\kappa\dot{c}^i_{\ell-n}\cdot} f\|_{\infty,\,{}^{I_{\ell-n}}D_\varrho} \\ &= e^{\gamma_\varrho}\|f\|_{\widehat{\infty}_{\ell-n,i},\,{}^{I_{\ell-n}}D_\varrho}. \end{aligned}$$

□

Nachdem im ersten Schritt der Reinterpolation die Ableitung eliminiert wurde, können die restlichen Reinterpolationsschritte leicht mit den Erkenntnissen aus der Betrachtung des Einfachschichtoperators abgeschätzt werden.

Satz 4.48 (Fehler der Reinterpolation)
Seien für $L \in \underline{p_{\mathcal{I}}}$ Sequenzen von Clustern $\mathfrak{s}_t^L$, $\mathfrak{s}_s^L$, deren dazugehörige Gebiete die Kontraktionseigenschaft (4.3.5) *erfüllen, und eine passende Sequenz an Richtungen $\mathfrak{s}_c^L$ gegeben. Weiter seien ein $\varrho \in \mathbb{R}_{>1}$ und $\sigma := \vartheta(\varrho)$ gegeben, der betrachtete Interpolationsoperator von Ordnung $m \in \mathbb{N}_0$ sei stabil nach Bemerkung 6 und es existiere ein $\gamma_\varrho \in \mathbb{R}_{>0}$, das* (4.3.10) *erfüllt. Dann gibt es zu jedem $q \in (\sigma^{-1}, 1)$ ein $c_{q,\varrho}$ wie in Lemma 4.24, mit dem für alle $i \in \underline{6}$, $\ell \in \underline{L}$ und für alle auf ${}^{I_{0,i}}D_\varrho$ holomorphen Funktionen f*

$$\begin{aligned} &\left\| \left(e^{\mathrm{i}\kappa\dot{c}^i_0\cdot} f - (\mathfrak{I}^{i,c_L}_{I_L} \circ \cdots \circ \mathfrak{I}^{i,c_1}_{I_1})[e^{\mathrm{i}\kappa\dot{c}^i_0\cdot} f] \right)' \right\|_{\infty, I_L} \\ &\quad \le \tfrac{2}{|I_{L,i}|} \tfrac{\omega_{I_{L,i},\varrho}}{a_\varrho - 1} \left((1 + e^{2\gamma_\varrho} c_{q,\varrho} q^m)^L - 1 \right) \|f\|_{\infty,\,{}^{I_{0,i}}D_\varrho} \end{aligned}$$

gilt, wobei $\omega_{I_{L,i},\varrho}$ wie in Lemma 4.46 definiert ist. Außerdem existieren zu jedem gegebenen $\widetilde{q} \in (q,1)$ ein $\mathcal{C}'_{\varrho,q} > 0$, welches von $c_{q,\varrho}$, der Wahl von Zwischengrößen in $(q,\widetilde{q})$, γ_ϱ und ϱ abhängt, so dass

$$m \ge \mathcal{C}'_{\varrho,q}(1 + \ln(L)) \quad \Rightarrow \quad \tfrac{2}{|I_{L,i}|} \tfrac{\omega_{I_{L,i},\varrho}}{a_\varrho - 1} \left((1 + e^{2\gamma_\varrho} c_{q,\varrho} q^m)^L - 1 \right) \le \tfrac{\omega_{I_{L,i},\varrho}}{|I_{L,i}|} \widetilde{q}^m$$

erfüllt ist.

Beweis: Seien ein $\ell \in \underline{L}$, $i \in \underline{6}$ und eine auf ${}^{I_{0,i}}D_\varrho$ holomorphe Funktion f gegeben. Erneut soll zuerst die Ableitung eliminiert werden. Die zu differenzierende Funktion ist als Komposition aus Polynomen, ebenen Wellen und einer holomorphen Funktion f auch auf der transformierten Bernstein-Ellipse ${}^{I_{L,i}}D_\varrho$ holomorph. Entsprechend kann mit demselben Vorgehen wie in Lemma 4.46 unter Verwendung der Cauchy-Integralformel

$$\begin{aligned} &\left\| \left(e^{\mathrm{i}\kappa\dot{c}^i_0\cdot} f - (\mathfrak{I}^{i,c_L}_{I_L} \circ \cdots \circ \mathfrak{I}^{i,c_1}_{I_1})[e^{\mathrm{i}\kappa\dot{c}^i_0\cdot} f] \right)' \right\|_{\infty, I_L} \\ &\quad \le \tfrac{2}{|I_{L,i}|} \tfrac{1}{a_\varrho - 1} \left\| e^{\mathrm{i}\kappa\dot{c}^i_0\cdot} f - (\mathfrak{I}^{i,c_L}_{I_L} \circ \cdots \circ \mathfrak{I}^{i,c_1}_{I_1})[e^{\mathrm{i}\kappa\dot{c}^i_0\cdot} f] \right\|_{\infty,\,{}^{I_L}D_\varrho} \\ &\quad \le \tfrac{2}{|I_{L,i}|} \tfrac{\omega_{I_{L,i},\varrho}}{a_\varrho - 1} \left\| e^{\mathrm{i}\kappa\dot{c}^i_0\cdot} f - (\mathfrak{I}^{i,c_L}_{I_L} \circ \cdots \circ \mathfrak{I}^{i,c_1}_{I_1})[e^{\mathrm{i}\kappa\dot{c}^i_0\cdot} f] \right\|_{\widehat{\infty}_{L,i},\,{}^{I_L}D_\varrho} \end{aligned}$$

gefolgert werden. Um weiter abschätzen zu können, wird erneut eine Teleskopsumme genutzt. Setze dazu wie im Fall des Einfachschichtoperators in Satz 4.27

$$\begin{aligned} T_1 :&= Id - \mathfrak{I}_{I_L}^{i,c_L} \circ \cdots \circ \mathfrak{I}_{I_1}^{i,c_1} \\ &= (Id - \mathfrak{I}_{I_1}^{i,c_1}) + (Id - \mathfrak{I}_{I_2}^{i,c_2}) \circ \mathfrak{I}_{I_1}^{i,c_1} + \cdots + (Id - \mathfrak{I}_{I_L}^{i,c_L}) \circ \mathfrak{I}_{I_{L-1}}^{i,c_{L-1}} \circ \cdots \circ \mathfrak{I}_{I_1}^{i,c_1}, \end{aligned}$$

was mit Lemma 4.47 zu

$$\begin{aligned} \left\| T_1[e^{\mathrm{i}\kappa\dot{c}_0^i \cdot} f] \right\|_{\widehat{\infty}_{L,i},\, {}^{I_L}D_\varrho} &= \left\| \sum_{\ell=1}^{L} (Id - \mathfrak{I}_{I_\ell}^{i,c_\ell})(\mathfrak{I}_{I_{\ell-1}}^{i,c_{\ell-1}} \circ \cdots \circ \mathfrak{I}_{I_1}^{i,c_1})[e^{\mathrm{i}\kappa\dot{c}_0^i \cdot} f] \right\|_{\widehat{\infty}_{L,i},\, {}^{I_L}D_\varrho} \\ &\overset{\Delta}{\leq} \sum_{\ell=1}^{L} e^{\gamma_\varrho} \left\| (Id - \mathfrak{I}_{I_\ell}^{i,c_\ell})(\mathfrak{I}_{I_{\ell-1}}^{i,c_{\ell-1}} \circ \cdots \circ \mathfrak{I}_{I_1}^{i,c_1})[e^{\mathrm{i}\kappa\dot{c}_0^i \cdot} f] \right\|_{\widehat{\infty}_{\ell,i},\, {}^{I_\ell}D_\varrho} \end{aligned}$$

führt. Die gewichtete Norm kann erneut in die ungewichtete umgeschrieben werden, wobei die ebene Welle vor dem Interpolationsoperator verschwindet. Anschließend kann die normale Interpolationsaussage auf Bernstein-Ellipsen aus Lemma 4.24 verwendet und mit Hilfe von (4.3.10) auf eine bekannte Form gebracht werden

$$\begin{aligned} &\sum_{\ell=1}^{L} e^{\gamma_\varrho} \left\| (Id - \mathfrak{I}_{I_\ell}^{i,c_\ell})(\mathfrak{I}_{I_{\ell-1}}^{i,c_{\ell-1}} \circ \cdots \circ \mathfrak{I}_{I_1}^{i,c_1})[e^{\mathrm{i}\kappa\dot{c}_0^i \cdot} f] \right\|_{\widehat{\infty}_{\ell,i},\, {}^{I_\ell}D_\varrho} \\ &\quad \leq \sum_{\ell=1}^{L} e^{\gamma_\varrho} c_{q,\varrho} q^m \left\| e^{-\mathrm{i}\kappa\dot{c}_\ell^i \cdot} \mathfrak{I}_{I_{\ell-1}}^{i,c_{\ell-1}} \circ \cdots \circ \mathfrak{I}_{I_1}^{i,c_1}[e^{\mathrm{i}\kappa\dot{c}_0^i \cdot} f] \right\|_{\infty,\, {}^{I_{\ell-1}}D_\varrho} \\ &\quad = \sum_{\ell=1}^{L} e^{\gamma_\varrho} c_{q,\varrho} q^m \left\| e^{-\mathrm{i}\kappa\dot{c}_\ell^i \cdot} e^{\mathrm{i}\kappa\dot{c}_0^i \cdot} e^{-\mathrm{i}\kappa\dot{c}_0^i \cdot} \mathfrak{I}_{I_{\ell-1}}^{i,c_{\ell-1}} \circ \cdots \circ \mathfrak{I}_{I_1}^{i,c_1}[e^{\mathrm{i}\kappa\dot{c}_0^i \cdot} f] \right\|_{\infty,\, {}^{I_{\ell-1}}D_\varrho} \\ &\quad \leq \sum_{\ell=1}^{L} e^{2\gamma_\varrho} c_{q,\varrho} q^m \left\| e^{-\mathrm{i}\kappa\dot{c}_0^i \cdot} \mathfrak{I}_{I_{\ell-1}}^{i,c_{\ell-1}} \circ \cdots \circ \mathfrak{I}_{I_1}^{i,c_1}[e^{\mathrm{i}\kappa\dot{c}_0^i \cdot} f] \right\|_{\infty,\, {}^{I_{\ell-1}}D_\varrho}. \end{aligned}$$

Hierauf kann das Lemma 4.26 angewendet werden, es folgt

$$\begin{aligned} &\sum_{\ell=1}^{L} e^{2\gamma_\varrho} c_{q,\varrho} q^m \left\| e^{-\mathrm{i}\kappa\dot{c}_0^i \cdot} \mathfrak{I}_{I_{\ell-1}}^{i,c_{\ell-1}} \circ \cdots \circ \mathfrak{I}_{I_1}^{i,c_1}[e^{\mathrm{i}\kappa\dot{c}_0^i \cdot} f] \right\|_{\infty,\, {}^{I_{\ell-1}}D_\varrho} \\ &\quad \leq e^{2\gamma_\varrho} c_{q,\varrho} q^m \sum_{\ell=1}^{L} (1 + e^{2\gamma_\varrho} c_{q,\varrho} q^m)^{\ell-1} \left\| e^{-\mathrm{i}\kappa\dot{c}_0^i \cdot} e^{\mathrm{i}\kappa\dot{c}_0^i \cdot} f \right\|_{\infty,\, {}^{I_{0,i}}D_\varrho} \\ &\quad = e^{2\gamma_\varrho} c_{q,\varrho} q^m \sum_{\ell=1}^{L} (1 + e^{2\gamma_\varrho} c_{q,\varrho} q^m)^{\ell-1} \left\| f \right\|_{\infty,\, {}^{I_{0,i}}D_\varrho}. \end{aligned}$$

Mit der Partialsumme der geometrischen Reihe ergibt sich dann

$$e^{2\gamma_\varrho} c_{q,\varrho} q^m \sum_{\ell=1}^{L} (1 + e^{2\gamma_\varrho} c_{q,\varrho} q^m)^{\ell-1} = (1 + e^{2\gamma_\varrho} c_{q,\varrho} q^m)^L - 1$$

und damit der erste Teil der Behauptung

$$\begin{aligned}&\left\| \left(e^{\mathrm{i}\kappa \dot{c}_0^i \cdot} f - (\mathfrak{I}_{I_L}^{i,c_L} \circ \cdots \circ \mathfrak{I}_{I_1}^{i,c_1})[e^{\mathrm{i}\kappa \dot{c}_0^i \cdot} f] \right)' \right\|_{\infty, I_L} \\ &\quad \leq \tfrac{2}{|I_{L,i}|} \tfrac{\omega_{I_{L,i},\varrho}}{a_\varrho - 1} \left((1 + e^{2\gamma_\varrho} c_{q,\varrho} q^m)^L - 1 \right) \|f\|_{\infty, {}^{I_{0,i}}D_\varrho}.\end{aligned}$$

Für den zweiten Teil der Behauptung kann ein Zwischenschritt aus dem Beweis des Satzes 4.27 genutzt werden. Dazu unterscheide zwei Fälle. Falls

$$\tfrac{2}{a_\varrho - 1} \leq 1$$

gilt, existiert zu jedem $\widetilde{q} \in (q, 1)$ und $\widehat{q} \in (q, \widetilde{q})$ nach Satz 4.27 eine Konstante $\mathcal{C}_{\varrho,q} > 0$ mit

$$m \geq \mathcal{C}_{\varrho,q}(1 + \ln(L)) \qquad \Rightarrow \qquad (1 + e^{2\gamma_\varrho} c_{q,\varrho} q^m)^L - 1 \leq \widetilde{q}^m.$$

Es folgt direkt

$$\tfrac{2}{|I_{L,i}|} \tfrac{\omega_{I_{L,i},\varrho}}{a_\varrho - 1} \left((1 + e^{2\gamma_\varrho} c_{q,\varrho} q^m)^L - 1 \right) \leq \tfrac{\omega_{I_{L,i},\varrho}}{|I_{L,i}|} (1 + e^{2\gamma_\varrho} c_{q,\varrho} q^m)^L - 1 \leq \tfrac{\omega_{I_{L,i},\varrho}}{|I_{L,i}|} \widetilde{q}^m$$

und damit die Behauptung für $\mathcal{C}'_{\varrho,q} := \mathcal{C}_{\varrho,q}$.
Ansonsten nutze ein $\overline{q} \in (q, \widetilde{q})$, um mit dem Satz 4.27 die Existenz eines $\mathcal{C}_{\varrho,q}$ zu erhalten, mit dem

$$m \geq \mathcal{C}_{\varrho,q}(1 + \ln(L)) \qquad \Rightarrow \qquad (1 + e^{2\gamma_\varrho} c_{q,\varrho} q^m)^L - 1 \leq (\overline{q})^m$$

gilt. Weiter nutze aus, dass $\frac{2}{a_\varrho - 1}$ unabhängig von m beschränkt ist, es existiert also zu dem $\widetilde{q}$ ein von ϱ abhängiges $m_\varrho \in \mathbb{N}$, so dass

$$\tfrac{2}{|I_{L,i}|} \tfrac{\omega_{I_{L,i},\varrho}}{a_\varrho - 1} \left((1 + e^{2\gamma_\varrho} c_{q,\varrho} q^m)^L - 1 \right) \leq \tfrac{2}{|I_{L,i}|} \tfrac{\omega_{I_{L,i},\varrho}}{a_\varrho - 1} (\overline{q})^m \leq \tfrac{\omega_{I_{L,i},\varrho}}{|I_{L,i}|} \widetilde{q}^m \quad \text{für alle } m > m_\varrho$$

gilt. Entsprechend passe $\mathcal{C}'_{\varrho,q}$ mit

$$\mathcal{C}'_{\varrho,q} := \max\{\mathcal{C}_{\varrho,q}, m_\varrho\}$$

an, um die Behauptung zu erhalten. □

Erneut kann mit den bisherigen Ergebnissen leicht eine Stabilitätskonstante für die Verkettung von Differenzial- und Interpolationsoperator $\frac{d}{dy_j} \circ \mathfrak{I}_I^{i,\mathfrak{s}_c^L}$ angegeben und eine Schranke für die Stabilitätskonstante bestimmt werden.

Lemma 4.49 (Stabilität der Reinterpolation)
Seien die Voraussetzungen von Satz 4.48 erfüllt. Dann existiert für feste q, ϱ und γ_ϱ eine Funktion $S^d_{q,\varrho,\gamma_\varrho}$, die für alle $L \in \underline{p_{\mathcal{I}}}, m \in \mathbb{N}_0$ durch

$$S^d_{q,\varrho,\gamma_\varrho}(L, m) = (1 + e^{2\gamma_\varrho} c_{q,\varrho} q^m)^{L-1} - 1 + e^{\gamma_\varrho}$$

gegeben ist. Mit der Funktion $S^d_{q,\varrho,\gamma_\varrho}$ gilt für alle $i \in \underline{6}_4$ mit $j = i - 3$ für die Operatornorm

$$\left\| \tfrac{d}{dy_j} \circ \mathfrak{I}_I^{i,\mathfrak{s}_c^L} \right\|_{op, C(I_L) \leftarrow C(I_0)} \leq \tfrac{2}{|I_{L,i}|} \tfrac{\omega_{I_{L,i},\varrho}}{a_\varrho - 1} \varrho^m \Lambda_m S^d_{q,\varrho,\gamma_\varrho}(L, m).$$

Beweis: Seien ein $i \in \underline{6}_{\lfloor 4}$ und eine Funktion $v_i \in C(I_{1,i})$ gegeben und definiere ein Polynom $p_{1,i} \in \Pi_m$ durch $p_{1,i} := \mathfrak{I}_{I_{1,i}}[e^{-\mathrm{i}\kappa\dot{c}_1^i \cdot} v_i]$. Dann gilt

$$\mathfrak{I}_{I_1}^{i,c_1}[v_i] = e^{\mathrm{i}\kappa\dot{c}_1^i \cdot} p_{1,i}.$$

Wie schon in Lemma 4.28 halte fest, dass sich dann

$$\|p_{1,i}\|_{\infty,{}^{I_1}D_\varrho} \leq \varrho^m \Lambda_m \|v_i\|_{\infty,I_0}$$

ergibt. Auch in diesem Fall kann mit einer Teleskopsumme die gesuchte Abschätzung auf den Satz 4.48 zurückgeführt werden. Nutze dabei nicht $e^{\mathrm{i}\kappa\dot{c}_0^i \cdot}$, sondern $e^{\mathrm{i}\kappa\dot{c}_1^i \cdot}$ als dauerhaft auftretende ebene Welle. Es folgt

$$\begin{aligned}
&\left\| \left(\mathfrak{I}_I^{i,\mathfrak{s}_c^L}[v_i] \right)' \right\|_{\infty,I_L} = \left\| \left(\mathfrak{I}_{I_L}^{i,c_L} \circ \cdots \circ \mathfrak{I}_{I_2}^{i,c_2}[e^{\mathrm{i}\kappa\dot{c}_1^i \cdot} p_{1,i}] \right)' \right\|_{\infty,I_L} \\
&= \left\| (e^{\mathrm{i}\kappa\dot{c}_1^i \cdot} p_{1,i})' - (e^{\mathrm{i}\kappa\dot{c}_1^i \cdot} p_{1,i})' + \left(\mathfrak{I}_{I_L}^{i,c_L} \circ \cdots \circ \mathfrak{I}_{I_2}^{i,c_2}[e^{\mathrm{i}\kappa\dot{c}_1^i \cdot} p_{1,i}] \right)' \right\|_{\infty,I_L} \\
&\overset{\Delta}{\leq} \|(e^{\mathrm{i}\kappa\dot{c}_1^i \cdot} p_{1,i})'\|_{\infty,I_L} + \left\| \left(e^{\mathrm{i}\kappa\dot{c}_1^i \cdot} p_{1,i} - \mathfrak{I}_{I_L}^{i,c_L} \circ \cdots \circ \mathfrak{I}_{I_2}^{i,c_2}[e^{\mathrm{i}\kappa\dot{c}_1^i \cdot} p_{1,i}] \right)' \right\|_{\infty,I_L} \\
&\overset{4.48}{\leq} \|(e^{\mathrm{i}\kappa\dot{c}_1^i \cdot} p_{1,i})'\|_{\infty,I_L} + \tfrac{2}{|I_{L,i}|} \tfrac{\omega_{I_{L,i},\varrho}}{a_\varrho - 1} \left((1 + e^{2\gamma_\varrho} c_{q,\varrho} q^m)^{L-1} - 1 \right) \|p_{1,i}\|_{\infty,{}^{I_1}D_\varrho}.
\end{aligned}$$

Mit der Cauchy-Integralformel kann die verbliebene Ableitung eliminiert werden, es ergibt sich

$$\begin{aligned}
\|(e^{\mathrm{i}\kappa\dot{c}_1^i \cdot} p_{1,i})'\|_{\infty,I_L} &\leq \tfrac{2}{|I_{L,i}|} \tfrac{1}{a_\varrho - 1} \|e^{\mathrm{i}\kappa\dot{c}_1^i \cdot} p_{1,i}\|_{\infty,{}^{I_L}D_\varrho} \\
&\leq \tfrac{2}{|I_{L,i}|} \tfrac{\omega_{I_{L,i},\varrho}}{a_\varrho - 1} \|e^{\mathrm{i}\kappa\dot{c}_1^i \cdot} e^{-\mathrm{i}\kappa\dot{c}_L^i \cdot} p_{1,i}\|_{\infty,{}^{I_L}D_\varrho} \leq \tfrac{2}{|I_{L,i}|} \tfrac{\omega_{I_{L,i},\varrho}}{a_\varrho - 1} e^{\gamma_\varrho} \|p_{1,i}\|_{\infty,{}^{I_L}D_\varrho} \\
&\leq \tfrac{2}{|I_{L,i}|} \tfrac{\omega_{I_{L,i},\varrho}}{a_\varrho - 1} e^{\gamma_\varrho} \|p_{1,i}\|_{\infty,{}^{I_1}D_\varrho}.
\end{aligned}$$

Insgesamt folgt

$$\begin{aligned}
&\left\| \left(\mathfrak{I}_I^{i,\mathfrak{s}_c^L}[v_i] \right)' \right\|_{\infty,I_L} \leq \tfrac{2}{|I_{L,i}|} \tfrac{\omega_{I_{L,i},\varrho}}{a_\varrho - 1} \left((1 + e^{2\gamma_\varrho} c_{q,\varrho} q^m)^{L-1} - 1 + e^{\gamma_\varrho} \right) \|p_{1,i}\|_{\infty,{}^{I_1}D_\varrho} \\
&\leq \tfrac{2}{|I_{L,i}|} \tfrac{\omega_{I_{L,i},\varrho}}{a_\varrho - 1} \left((1 + e^{2\gamma_\varrho} c_{q,\varrho} q^m)^{L-1} - 1 + e^{\gamma_\varrho} \right) \varrho^m \Lambda_m \|v_i\|_{\infty,I_0}
\end{aligned}$$

und zusammen mit der Funktion $S_{q,\varrho,\gamma_\varrho}^d$ die Behauptung zur Operatornorm. □

Zum Schluss folgt eine Abschätzung der Operatornorm im Mehrdimensionalen.

Theorem 4.50 (Stabilität mehrdimensionale Reinterpolation)
Seien für $L \in \underline{p_{\mathcal{I}}}_{\rfloor}$ Sequenzen an Clustern $\mathfrak{s}_t^L$, $\mathfrak{s}_s^L$, deren dazugehörige Gebiete die Kontraktionseigenschaft (4.3.5) *erfüllen, und eine passende Sequenz an Richtungen $\mathfrak{s}_c^L$ gegeben. Weiter seien ein $\varrho \in \mathbb{R}_{>1}$ und $\sigma := \vartheta(\varrho)$ gegeben, es existiere ein $\gamma_\varrho \in \mathbb{R}_{>0}$, das* (4.3.10)

erfüllt, und der betrachtete Interpolationsoperator von Ordnung $m \in \mathbb{N}_0$ sei stabil nach Bemerkung 6. Dann gilt für jedes $q \in (\sigma^{-1}, 1)$, $p \in (q, 1]$ und jede Interpolationsordnung

$$m \geq \frac{\ln(L)}{\ln\left(\frac{p}{q}\right)},$$

mit der Funktion $S^d_{q,\varrho,\gamma_\varrho}$ aus Lemma 4.49 für die Operatornorm der Reinterpolation im Mehrdimensionalen für alle $j \in \underline{3}_|$ mit $j' = j + 3$

$$\left\| \tfrac{\partial}{\partial y_j} \circ \left(\mathfrak{I}^{c_L}_{Q_{t_L} \times Q_{s_L}} \circ \cdots \circ \mathfrak{I}^{c_1}_{Q_{t_1} \times Q_{s_1}} \right) \right\|_{op, C(Q_{t_L} \times Q_{s_L}) \leftarrow C(Q_{t_0} \times Q_{s_0})} \leq \tfrac{2}{|I_{L,j'}|} \tfrac{\omega_{I_{L,j'},\varrho}}{a_\varrho - 1} \varrho^m \Lambda_m^6$$
$$\cdot (p^m c_{q,\varrho} \widehat{c}_{\gamma_\varrho,q,\varrho} + 1)^5 S^d_{q,\varrho,\gamma_\varrho}(L, m),$$

wobei $\widehat{c}_{\gamma_\varrho,q,\varrho}$ die Konstante aus Lemma 4.28 ist.

Beweis: Da es sich hier ebenfalls um einen Tensorinterpolationsoperator handelt, können die einzelnen Interpolationsoperatoren umsortiert werden. Weiterhin kann die partielle Ableitung an unbeteiligten Interpolationsoperatoren vorbeigezogen werden. Für $j \in \underline{3}_|$ und $j' = j + 3$ ergibt sich die Behauptung aus der Kombination von Lemma 4.28 und Lemma 4.49 mit

$$\begin{aligned}
&\left\| \tfrac{\partial}{\partial y_j} \circ \left(\mathfrak{I}^{c_L}_{Q_{t_L} \times Q_{s_L}} \circ \cdots \circ \mathfrak{I}^{c_1}_{Q_{t_1} \times Q_{s_1}} \right) \right\|_{op, C(Q_{t_L} \times Q_{s_L}) \leftarrow C(Q_{t_0} \times Q_{s_0})} \\
&\leq \left\| \tfrac{d}{dy_j} \circ \mathfrak{I}^{j', \mathfrak{s}^L_c}_I \right\|_{op, C(I_L) \leftarrow C(I_0)} \prod_{\substack{i=1 \\ i \neq j'}}^{6} \left\| \mathfrak{I}^{i, \mathfrak{s}^L_c}_I \right\|_{op, C(I_L) \leftarrow C(I_0)} \\
&\leq \tfrac{2}{|I_{L,j'}|} \tfrac{\omega_{I_{L,j'},\varrho}}{a_\varrho - 1} \varrho^m \Lambda_m^6 (p^m c_{q,\varrho} \widehat{c}_{\gamma_\varrho,q,\varrho} + 1)^5 S^d_{q,\varrho,\gamma_\varrho}(L, m).
\end{aligned}$$

□

Die Stabilitätskonstante beim Doppelschichtoperator ist erkennbar schlechter als die beim Einfachschichtoperator, glücklicherweise kommt der Term ϱ^m jedoch nur in einer Dimension vor.
Auch die Ergebnisse zum Reinterpolationsfehler lassen sich direkt auf den adjungierten Doppelschichtoperator übertragen.

4.5 Numerische Experimente

Nach der aufwendigen Bestimmung von Schranken für die Approximationsfehler sollen ein paar Experimente zum tatsächlich auftretenden Fehler durchgeführt werden. Es werden die Wellenzahl (κ) sowie die Interpolationsordnung (m) variiert. Die Zulässigkeitsparameter

m	κ	$\|A_e - \widetilde{A}_e\|_2$	$\frac{\|A_e - \widetilde{A}_e\|_2}{\|A_e\|_2}$	$\|A_e - \widetilde{A}_e\|_F$	κ	$\|A_e - \widetilde{A}_e\|_2$	$\frac{\|A_e - \widetilde{A}_e\|_2}{\|A_e\|_2}$	$\|A_e - \widetilde{A}_e\|_F$
0	8	1.54_{-5}	1.65_{-1}	1.31_{-4}	32	2.95_{-6}	8.32_{-2}	5.73_{-5}
1	8	6.89_{-6}	7.40_{-2}	3.70_{-5}	32	1.39_{-6}	3.93_{-2}	1.60_{-5}
2	8	4.29_{-7}	4.61_{-3}	2.96_{-6}	32	1.05_{-7}	2.96_{-3}	1.60_{-6}
3	8	4.87_{-8}	5.23_{-4}	2.93_{-7}	32	1.22_{-8}	3.45_{-4}	1.69_{-7}
4	8	3.91_{-9}	4.19_{-5}	2.29_{-8}	32	1.04_{-9}	2.95_{-5}	1.63_{-8}
5	8	3.10_{-10}	3.33_{-6}	1.63_{-9}	32	1.13_{-10}	3.20_{-6}	1.40_{-9}
6	8	2.09_{-11}	2.24_{-7}	1.02_{-10}	32	7.81_{-12}	2.21_{-7}	1.08_{-10}

Tabelle 4.1: Fehler der Approximation des Einfachschichtoperators bei verschiedenen Interpolationsordnungen auf der Sphäre ($n = 32768$)

werden mit $\eta_1 = 10$ und $\eta_2 = 1$ gewählt und als Grundgeometrie die Einheitssphäre sowie ein Würfel ($[-1, 1]^3$) mit unterschiedlichen Problemgrößen ($n := \#\mathcal{I}$) verwendet.
Alle Rechnungen wurden auf einem *Shared Memory* System mit zwei Intel® Xeon® Platinum 8160 Prozessoren mit insgesamt 48 Kernen durchgeführt.
Zu Beginn betrachte den Einfachschichtoperator. Um eine möglichst große Spanne an Interpolationsordnungen abdecken zu können, nutze eine kleine Problemgröße ($n = 32768$) und starte bei $m = 0$, also einem Interpolationspunkt pro Dimension. Der Fehler wird für die gesamte Matrix gemessen, dazu werden approximativ die Spektralnorm ($\|A_e - \widetilde{A}_e\|_2$) und die exakte Frobeniusnorm ($\|A_e - \widetilde{A}_e\|_F$) bestimmt. Die Ergebnisse für einen niedrigfrequenten ($\kappa = 8$) sowie einen hochfrequenten ($\kappa = 32$) Einfachschichtoperator auf der Sphäre finden sich in Tabelle 4.1.

Ähnliche Ergebnisse liefert die Betrachtung des Einfachschichtoperators auf dem Würfel (siehe Tabelle 4.2), dieses Mal mit $n = 37632$ und Wellenzahlen $\kappa = 3.5$ und $\kappa = 14$ für den niedrig- und hochfrequenten Fall.

In den Tabellen 4.1 und 4.2 finden sich die relativen Fehler in der Spektralnorm, in der Abbildung 4.4 der relative Fehler in der Frobeniusnorm bei Veränderung der Interpolationsordnung m. Es zeigt sich, dass alle Problemstellungen eine ähnliche Konvergenzrate aufweisen. Teilweise produziert der hochfrequente Fall kleinere Fehler, was daran liegt, dass beim gleichen Experiment mit höheren Wellenzahlen die Zulässigkeitsbedingungen schärfer und daher Blöcke erst später zulässig werden.
Die nach der Theorie erwartete exponentielle Konvergenz ist für den hochfrequenten Fall noch einmal grafisch in Abbildung 4.5 zu finden. Die Konstante C wurde dabei so gewählt, dass die Exponentialfunktion ungefähr mittig zwischen den beiden Fehlern liegt.

m	κ	$\|A_e - \widetilde{A}_e\|_2$	$\frac{\|A_e - \widetilde{A}_e\|_2}{\|A_e\|_2}$	$\|A_e - \widetilde{A}_e\|_F$	κ	$\|A_e - \widetilde{A}_e\|_2$	$\frac{\|A_e - \widetilde{A}_e\|_2}{\|A_e\|_2}$	$\|A_e - \widetilde{A}_e\|_F$
0	3.5	7.93_{-5}	2.70_{-1}	3.21_{-4}	14	5.15_{-5}	5.08_{-1}	3.65_{-4}
1	3.5	8.60_{-6}	2.93_{-2}	3.55_{-5}	14	1.08_{-5}	1.06_{-1}	9.70_{-5}
2	3.5	8.86_{-7}	3.02_{-3}	3.24_{-6}	14	2.42_{-6}	2.39_{-2}	1.69_{-5}
3	3.5	8.49_{-8}	2.89_{-4}	2.72_{-7}	14	4.73_{-7}	4.67_{-3}	2.67_{-6}
4	3.5	6.73_{-9}	2.29_{-5}	2.09_{-8}	14	7.42_{-8}	7.32_{-4}	3.81_{-7}
5	3.5	4.66_{-10}	1.59_{-6}	1.44_{-9}	14	1.03_{-8}	1.01_{-4}	4.93_{-8}
6	3.5	2.73_{-11}	9.32_{-8}	9.48_{-11}	14	1.20_{-9}	1.18_{-5}	5.60_{-9}

Tabelle 4.2: Fehler der Approximation des Einfachschichtoperators bei verschiedenen Interpolationsordnungen auf dem Würfel ($n = 37632$)

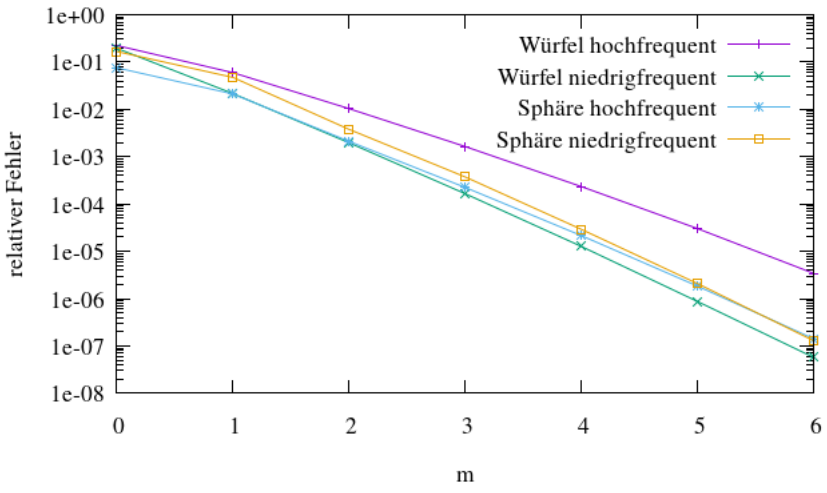


Abbildung 4.4: Vergleich der Konvergenz des Einfachschichtoperators gemessen in der relativen Frobeniusnorm auf der Sphäre und dem Würfel

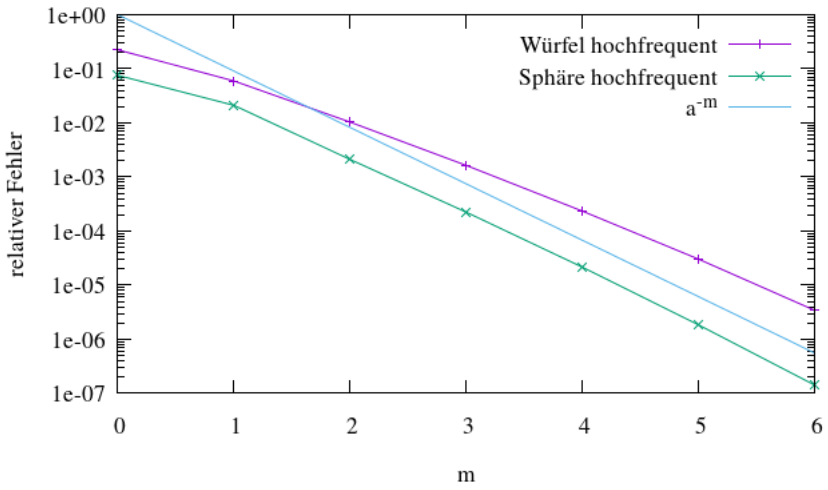


Abbildung 4.5: Exponentielle Konvergenz beim Einfachschichtoperator

Ebendiese Experimente wurden noch einmal für den Doppelschichtoperator wiederholt. Auch hier wird zunächst die Sphäre bei wachsender Interpolationsordnung und fester Problemgröße $n = 32768$ betrachtet. Die Ergebnisse finden sich in Tabelle 4.3.

Bei den Ergebnissen des Doppelschichtoperators ist gut zu erkennen, dass die Konvergenz sich gegenüber der des Einfachschichtoperators verschlechtert. Die nach der Theorie erwartete Verschlechterung der Konvergenzrate zeigt sich in der Tabelle 4.3 durch Fehler, die beim gleichen Experiment für den Einfachschichtoperator (siehe Tabelle 4.1) ungefähr um den Faktor 10 kleiner sind. Diese Beobachtungen bezüglich der Konvergenz setzen sich auch für den Würfel als Geometrie fort. Die entsprechenden Ergebnisse für den Würfel finden sich in der Tabelle 4.4.

Die grafischen Auswertungen der Ergebnisse ähneln denen vom Einfachschichtoperatorfall. In der Abbildung 4.6 sind erneut die relativen Fehler der einzelnen Problemstellungen in der Frobeniusnorm und im Vergleich dargestellt. Erstaunlicherweise sind die Ergebnisse der Sphäre sowohl im hochfrequenten als auch im niedrigfrequenten Fall näher beieinander als beim Einfachschichtoperator. Dafür zeigt sich auch hier sehr deutlich der Qualitätsunterschied beim hoch- und niedrigfrequenten Würfel.
Die exponentielle Konvergenz im Fall des hochfrequenten Doppelschichtoperators ist noch einmal grafisch in Abbildung 4.7 dargestellt, auch hier wurde die Vergleichskurve mit Hilfe einer Konstante C so gelegt, dass sie zwischen den beiden Kurven aus den Experimenten liegt.

Die durchgeführten Experimente bestätigen die Aussagen der Theorie zur Konvergenz be-

m	κ	$\|A_d - \widetilde{A}_d\|_2$	$\frac{\|A_d - \widetilde{A}_d\|_2}{\|A_d\|_2}$	$\|A_d - \widetilde{A}_d\|_F$	κ	$\|A_d - \widetilde{A}_d\|_2$	$\frac{\|A_d - \widetilde{A}_d\|_2}{\|A_d\|_2}$	$\|A_d - \widetilde{A}_d\|_F$
0	8	3.00_{-4}	3.82_{-1}	1.62_{-3}	32	1.31_{-4}	1.65_{-1}	1.66_{-3}
1	8	5.63_{-5}	7.16_{-2}	3.42_{-4}	32	4.04_{-5}	5.08_{-2}	4.44_{-4}
2	8	1.00_{-5}	1.27_{-2}	4.94_{-5}	32	3.35_{-6}	4.21_{-3}	4.90_{-5}
3	8	1.14_{-6}	1.45_{-3}	5.23_{-6}	32	4.72_{-7}	5.93_{-4}	5.51_{-6}
4	8	1.28_{-7}	1.63_{-4}	4.98_{-7}	32	4.79_{-8}	6.02_{-5}	5.65_{-7}
5	8	1.11_{-8}	1.41_{-5}	3.95_{-8}	32	6.09_{-9}	7.65_{-6}	5.34_{-8}
6	8	8.67_{-10}	1.10_{-6}	2.82_{-9}	32	5.45_{-10}	6.84_{-7}	4.57_{-9}

Tabelle 4.3: Fehler der Approximation des Doppelschichtoperators bei verschiedenen Interpolationsordnungen auf der Sphäre ($n = 32768$)

m	κ	$\|A_d - \widetilde{A}_d\|_2$	$\frac{\|A_d - \widetilde{A}_d\|_2}{\|A_d\|_2}$	$\|A_d - \widetilde{A}_d\|_F$	κ	$\|A_d - \widetilde{A}_d\|_2$	$\frac{\|A_d - \widetilde{A}_d\|_2}{\|A_d\|_2}$	$\|A_d - \widetilde{A}_d\|_F$
0	3.5	4.24_{-4}	4.28_{-1}	1.65_{-3}	14	3.95_{-4}	3.87_{-1}	4.10_{-3}
1	3.5	1.01_{-4}	1.02_{-1}	3.84_{-4}	14	2.29_{-4}	2.24_{-1}	1.56_{-3}
2	3.5	2.05_{-5}	2.07_{-2}	5.97_{-5}	14	1.04_{-4}	1.01_{-1}	5.13_{-4}
3	3.5	3.28_{-6}	3.30_{-3}	7.42_{-6}	14	3.36_{-5}	3.28_{-2}	1.31_{-4}
4	3.5	4.00_{-7}	4.03_{-4}	7.94_{-7}	14	8.47_{-6}	8.29_{-3}	2.78_{-5}
5	3.5	3.94_{-8}	3.97_{-5}	7.56_{-8}	14	1.68_{-6}	1.64_{-3}	5.10_{-6}
6	3.5	3.19_{-9}	3.21_{-6}	7.35_{-9}	14	2.71_{-7}	2.65_{-4}	8.04_{-7}

Tabelle 4.4: Fehler der Approximation des Doppelschichtoperators bei verschiedenen Interpolationsordnungen auf dem Würfel ($n = 37632$)

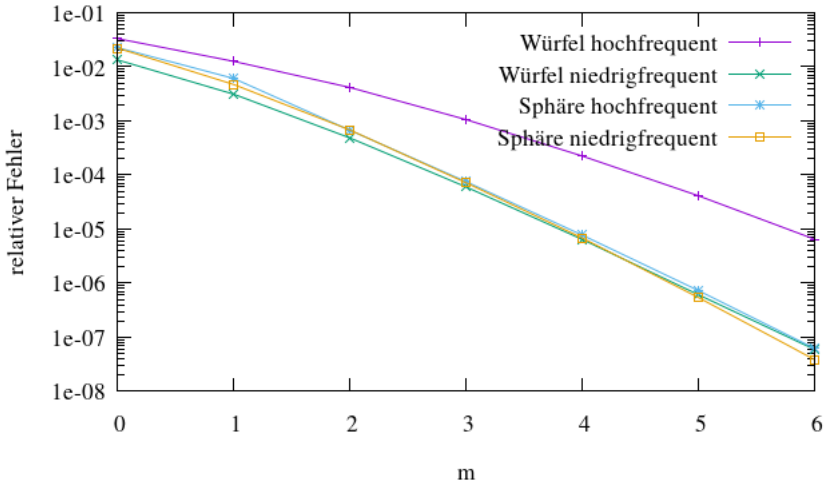


Abbildung 4.6: Vergleich der Konvergenz des Doppelschichtoperators auf der Sphäre und dem Würfel

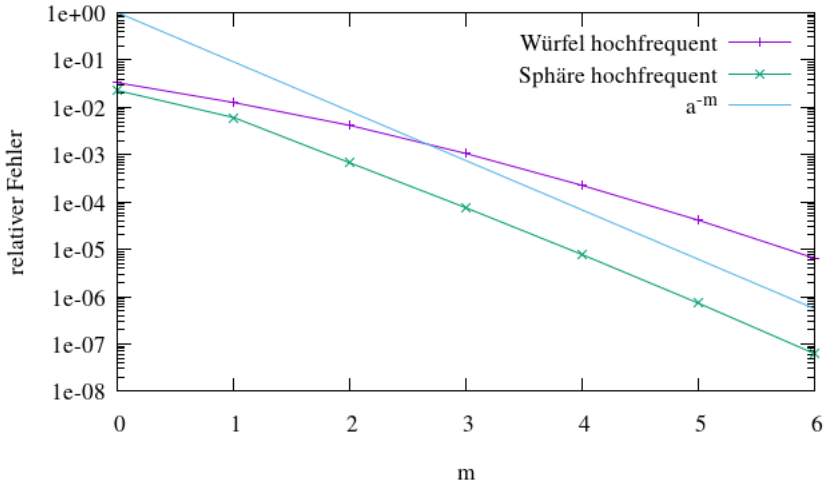


Abbildung 4.7: Exponentielle Konvergenz beim Doppelschichtoperator

züglich der beiden untersuchten Potentiale. Da die Konvergenzrate entscheidend von der Interpolationsordnung m abhängt, der Rang der Basen und Kopplungsmatrizen je nach Wahl von k aber mit m^3 wächst, sollte noch untersucht werden, wie der Rang weiter reduziert werden kann. Dies ist Thema des anschließenden Kapitels.

5 Aufstellen und Komprimieren von $\mathcal{RH}^2$-Matrizen

Neben der schon erläuterten direkten Berechnung der $\mathcal{RH}^2$-Matrizen mit Hilfe der Tensorinterpolation gibt es auch andere Varianten, eine Matrix dieses Typs aufzustellen, beziehungsweise schon vorhandene $\mathcal{RH}^2$-Matrizen-Darstellungen effizienter zu gestalten. In diesem Kapitel werden eine alternative Methode zum Generieren der Matrix sowie zwei Möglichkeiten, eine bereits gegebene $\mathcal{RH}^2$-Matrizen-Darstellung zu verbessern, vorgestellt und analysiert.
Ein wichtiges Hilfsmittel bei all diesen Methoden sind normerhaltende Matrizen.

Definition 5.1 (Unitäre und isometrische Matrizen)
Sei für $M, N \subset \mathbb{N}$ eine Matrix $Q \in \mathbb{C}^{M\times N}$ gegeben. Falls die Matrix $Q^*Q = Id$ erfüllt, bezeichne Q als isometrisch*, ist zusätzlich noch $QQ^* = Id$ erfüllt, bezeichne Q als* unitär.

Die Bezeichnung *isometrisch* leitet sich aus der Eigenschaft, normerhaltend zu sein, ab und ist eine Verallgemeinerung der Klasse der unitären Matrizen auf rechteckige Formate. Seien für $M, N \subset \mathbb{N}$ eine isometrische Matrix $Q \in \mathbb{C}^{M\times N}$ sowie eine Matrix $A \in \mathbb{C}^{N\times M}$ und ein Vektor $x \in \mathbb{C}^M$ gegeben. Dann gilt

$$\|QAx\|_2^2 = \langle QAx, QAx\rangle_2 = \langle Ax, Q^*QAx\rangle_2 = \langle Ax, Ax\rangle_2 = \|Ax\|_2^2,$$

ebenso folgt über die Eigenschaften der Spektralnorm $\|A^*Q^*\|_2 = \|A^*\|_2$.
Beides lässt sich auch auf die Frobeniusnorm übertragen, es gilt also $\|QA\|_F = \|A\|_F$ sowie $\|A^*Q^*\|_F = \|A^*\|_F$.

5.1 Orthogonalisierung richtungsabhängiger Clusterbasen

Eine Clusterbasis mit isometrischen Matrizen V_{tc} bietet bei vielen Algorithmen Vorteile und kann teilweise auch die Ränge reduzieren. Da die Ränge sowohl beim Speichern als auch bei der Arithmetik in die Komplexität eingehen, sind kleinere und insbesondere nicht unnötig hohe Ränge für die Praxis erstrebenswert.

Definition 5.2 (Isometrische richtungsabhängige Clusterbasis)
Bezeichne eine richtungsabhängige Clusterbasis $\{V_{tc}\}_{\substack{t\in\mathcal{T}_\mathcal{I}\\ c\in\mathcal{R}_t}}$ *als* isometrisch*, falls*

$$V_{tc}^* V_{tc} = Id \qquad \text{für alle } t \in \mathcal{T}_\mathcal{I}, c \in \mathcal{R}_t \text{ mit } k_{tc} \neq 0$$

gilt.

Für die Anwendung ist die Definition allein noch zu unhandlich. Um eine isometrische richtungsabhängige Clusterbasis effizient aufstellen zu können, werden Eigenschaften aller zugehöriger Matrizen eingeführt.
Zunächst jedoch noch eine kurze Bemerkung vorweg. Betrachte ein $t \in \mathcal{T}_\mathcal{I}$ mit zwei Kindern $t_1, t_2 \in \operatorname{kind}(t), t_1 \neq t_2$, dann gilt für alle Richtungen $c \in \mathcal{R}_t$ mit $c' = \mathrm{r}_t(c)$

$$V_{t_1c'}^* V_{t_2c'} = 0.$$

Denn aus $t_1 \neq t_2$ folgt nach Definition 2.13 ${}^\mathfrak{I}t_1 \cap {}^\mathfrak{I}t_2 = \emptyset$ und entsprechend ergibt sich aus $V_{t_1c'} \in \mathbb{C}^{\mathcal{I}\times k_{t_1c'}}_{{}^\mathfrak{I}t_1\times k_{t_1c'}}$ und $V_{t_2c'} \in \mathbb{C}^{\mathcal{I}\times k_{t_2c'}}_{{}^\mathfrak{I}t_2\times k_{t_2c'}}$, dass das Produkt der beiden Matrizen nur Nullen als Einträge hat. Dies ist noch einmal im folgenden Lemma festgehalten.

Lemma 5.3
Für alle Cluster $t \in \mathcal{T}_\mathcal{I}^\ell$ *mit* $\ell \in \underline{p_\mathcal{I}}$ *und Richtungen* $c \in \mathcal{R}_\ell$, $c' = \mathrm{r}_\ell(c) \in \mathcal{R}_{\ell+1}$ *einer richtungsabhängigen Clusterbasis* $\{V_{tc}\}_{\substack{t\in\mathcal{T}_\mathcal{I}\\ c\in\mathcal{R}_t}}$ *gilt*

$$V_{t_1c'}^* V_{t_2c'} = 0 \qquad \text{für alle } t_1, t_2 \in \operatorname{kind}(t) \text{ mit } t_1 \neq t_2.$$

Beweis: Siehe oben. □

Der Beweis für die Bedingungen der Orthogonalität für richtungsabhängige Clusterbasen orientiert sich an der Variante für $\mathcal{H}^2$-Matrizen in [4, Lem. 5.3].

Lemma 5.4 (Bedingung Orthogonalität)
Eine richtungsabhängige Clusterbasis $\{V_{tc}\}_{\substack{t\in\mathcal{T}_\mathcal{I}\\ c\in\mathcal{R}_t}}$ *ist genau dann isometrisch, wenn*

$$V_{tc}^* V_{tc} = Id \qquad \text{für alle } t \in \mathcal{L}_\mathcal{I}, c \in \mathcal{R}_t \text{ mit } k_{tc} \neq 0, \tag{5.1.1}$$

$$\sum_{t'\in\operatorname{kind}(t)} E_{t'c}^* E_{t'c} = Id \qquad \text{für alle } t \in \mathcal{T}_\mathcal{I} \setminus \mathcal{L}_\mathcal{I}, c \in \mathcal{R}_t \text{ mit } k_{tc} \neq 0 \tag{5.1.2}$$

gilt.

Beweis: "⇒" Nehme zunächst an, dass die richtungsabhängige Clusterbasis isometrisch ist. Seien $t \in \mathcal{T}_\mathcal{I}$ und ein $c \in \mathcal{R}_t$ mit $k_{tc} \neq 0$ gegeben. Der Fall $\operatorname{kind}(t) = \emptyset$ ist trivial, da die Definition einer isometrischen Clusterbasis schon die gewünscht Aussage liefert.

Deswegen sei im Folgenden $\mathrm{kind}(t) \neq \emptyset$. Betrachte die Summe der Transfermatrizen zu den Kindern von t mit $c' = \mathrm{r}_t(c)$ und nutze die Annahme, dass die richtungsabhängige Clusterbasis isometrisch ist

$$\sum_{t' \in \mathrm{kind}(t)} E_{t'c}^* E_{t'c} = \sum_{t' \in \mathrm{kind}(t)} E_{t'c}^* V_{t'c'}^* V_{t'c'} E_{t'c} \overset{5.3}{=} \sum_{t_1 \in \mathrm{kind}(t)} \sum_{t_2 \in \mathrm{kind}(t)} E_{t_1c}^* V_{t_1c'}^* V_{t_2c'} E_{t_2c}$$
$$= \left(\sum_{t_1 \in \mathrm{kind}(t)} V_{t_1c'} E_{t_1c} \right)^* \left(\sum_{t_2 \in \mathrm{kind}(t)} V_{t_2c'} E_{t_2c} \right) = V_{tc}^* V_{tc} = Id.$$

Die Behauptung folgt direkt mit der Definition 3.2 der richtungsabhängigen Clusterbasis.
"$\Leftarrow$" Für die andere Richtung nehme an, dass die Bedingungen für die Matrizen in den Blättern (5.1.1) und Transfermatrizen (5.1.2) gelten. Zeige die Behauptung per abschnittsweiser Induktion über die Kardinalität der Menge der Nachfahren $\#\,\mathrm{nac}(t)$.
I.A. Sei $\#\,\mathrm{nac}(t) = 1$, dann folgt die Behauptung direkt aus (5.1.1).
I.V. Sei $n \in \mathbb{N}$ so gegeben, dass die Behauptung für alle t mit $\#\,\mathrm{nac}(t) \leq n$ gilt.
I.S. Sei ein $t \in \mathcal{T}_{\mathcal{I}}$ mit $\#\,\mathrm{nac}(t) = n + 1$ sowie ein $c \in \mathcal{R}_t$ mit $k_{tc} \neq 0$ gegeben. Weiter sei $t' \in \mathrm{kind}(t)$, dann gilt $\#\,\mathrm{nac}(t') \leq \#\,\mathrm{nac}(t) - 1 = n$ und entsprechend für $c' = \mathrm{r}_t(c)$ nach der Induktionsvoraussetzung

$$V_{t'c'}^* V_{t'c'} = Id.$$

Somit ergibt sich

$$V_{tc}^* V_{tc} = \left(\sum_{t_1 \in \mathrm{kind}(t)} V_{t_1c'} E_{t_1c} \right)^* \left(\sum_{t_2 \in \mathrm{kind}(t)} V_{t_2c'} E_{t_2c} \right)$$
$$= \sum_{t' \in \mathrm{kind}(t)} E_{t'c}^* V_{t'c'}^* V_{t'c'} E_{t'c} \overset{I.V.}{=} \sum_{t' \in \mathrm{kind}(t)} E_{t'c}^* E_{t'c} \overset{(5.1.2)}{=} Id$$

und damit die Behauptung. □

Um eine schon vorhandene richtungsabhängige Clusterbasis zu einer isometrischen zu machen, während die geschachtelte Struktur erhalten bleibt, wird ein Basiswechsel vorgenommen. Die Konstruktion der neuen Basen geschieht unter Zuhilfenahme der Bedingungen aus Lemma 5.4, hilfreich ist dabei eine *QR-Zerlegung* [32, S. 33].

Definition und Lemma 5.5 (QR-Zerlegung)
Zu jeder Matrix $A \in \mathbb{C}^{M \times N}$ mit $p = \min\{\#M, \#N\}$ existieren eine Indexmenge $P \subset M$ mit $p = \#P$ sowie eine isometrische Matrix $Q \in \mathbb{C}^{M \times P}$ und eine obere Dreiecksmatrix $R \in \mathbb{C}^{P \times N}$ (für eine beliebige Ordnung der Indices), die

$$A = QR$$

erfüllen. Bezeichne das Paar (Q, R) als reduzierte QR-Zerlegung *von A.*

Eine Möglichkeit, die Voraussetzungen einer isometrischen Clusterbasis zu erfüllen, ist, eine reduzierte QR-Zerlegung der beteiligten Matrizen durchzuführen. Eine numerisch stabile Möglichkeit, diese zu berechnen, stellen Householder-Spiegelungen[i] dar. Householder-Spiegelungen ermöglichen es, einen Vektor auf ein Vielfaches eines anderen Vektors (in beiden Fällen ist der Null-Vektor ausgenommen) abzubilden. In der Praxis wird auf ein Vielfaches des ersten Einheitsvektors abgebildet. Wird dies rekursiv von der ersten Spalte der Matrix als abzubildenden Vektor beginnend durchgeführt, lässt sich eine obere Dreiecksmatrix gewinnen. Das Produkt der einzelnen Householder-Spiegelungen liefert die unitäre Matrix. Für nähere Information zur QR-Zerlegung und der Householder-Spiegelung sei der Leser zum Beispiel auf [26, §5] verwiesen.
Bleibt die Frage, wie ein Algorithmus zum Orthogonalisieren einer Clusterbasis aussieht. Seien ein Cluster $t \in \mathcal{T}_\mathcal{I}$ und eine Richtung $c \in \mathcal{R}_t$ mit $k_{tc} \neq 0$ gegeben. Falls t ein Blattcluster ist, ist das Ziel direkt mit einer reduzierten QR-Zerlegung der Matrix V_{tc} erreicht. Entsprechend gibt es für die Matrix $V_{tc} \in \mathbb{C}^{\mathcal{I} \times k_{tc}}_{\mathfrak{I}t \times k_{tc}}$ ein $p_{tc} = \min\left\{\#^\mathfrak{I} t, k_{tc}\right\}$ sowie eine isometrische Matrix $Q_{tc} \in \mathbb{C}^{\mathcal{I} \times p_{tc}}_{\mathfrak{I}t \times p_{tc}}$ und eine obere Dreiecksmatrix $R_{tc} \in \mathbb{C}^{p_{tc} \times k_{tc}}$ mit

$$V_{tc} = Q_{tc} R_{tc}.$$

Ersetzen der ursprünglichen Matrix V_{tc} durch $V_{tc}^{new} := Q_{tc}$ liefert eine isometrische Matrix der Clusterbasis, die obere Dreiecksmatrix R_{tc} beschreibt den Basiswechsel.

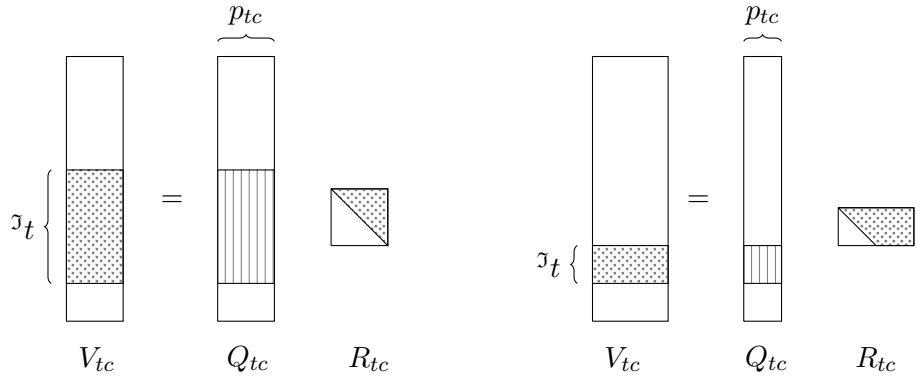


Abbildung 5.1: QR-Zerlegungen von Blattmatrizen

Die Abbildung 5.1 zeigt schraffiert die Anteile der beteiligten Matrizen, die von null verschieden sind. Im rechten Teil der Abbildung ist der Fall $k_{tc} > \#^\mathfrak{I} t$ gezeigt, bei dem eine Reduktion des Rangs auftritt. Gerade bei einem hohen Rang k_{tc}, zum Beispiel bei einer hohen Interpolationsordnung im Vergleich zu der Anzahl der Elemente in den Blättern, kann

[i] Diese Art der Transformation wurde 1958 in einem Artikel von dem amerikanischen Mathematiker Alston Scott Householder verwendet, um die QR-Zerlegung effizienter zu berechnen [35].

die Rangreduktion deutlich sein.
Handelt es sich bei dem betrachteten Cluster um kein Blattcluster, wird die Aufgabe anspruchsvoller, denn die geschachtelte Struktur der Clusterbasis soll durch die Orthogonalisierung nicht zerstört werden.
Sei $t \in \mathcal{T}_\mathcal{I} \setminus \mathcal{L}_\mathcal{I}$ so gegeben, dass für seine Kinder die Orthogonalisierung schon durchgeführt wurde. Für jede Richtung $c \in \mathcal{R}_t$ mit $k_{tc} \neq 0$ lässt sich die Matrix V_{tc} mit Hilfe ihrer Kinder $t' \in \operatorname{kind}(t)$ und der dazugehörigen Richtung $c' = \mathrm{r}_t(c) \in \mathcal{R}_{t'}$ über die Matrizen der Kinder $V_{t'c'}$ und die Transfermatrizen $E_{t'c}$ darstellen

$$V_{tc} = \sum_{t' \in \operatorname{kind}(t)} V_{t'c'} E_{t'c}.$$

Sei $\operatorname{kind}(t) = \{t'_1, \dots t'_\tau\}$ und setze $q_{tc} := \sum_{i=1}^{\tau} p_{t'_i c'}$. Einsetzen der QR-Zerlegungen der Kinder führt zu folgender Schreibweise der Clusterbasis

$$V_{tc} = \sum_{t' \in \operatorname{kind}(t)} Q_{t'c'} R_{t'c'} E_{t'c} = \underbrace{\begin{pmatrix} Q_{t'_1 c'} & \cdots & Q_{t'_\tau c'} \end{pmatrix}}_{=: U_{tc} \in \mathbb{C}^{\mathcal{I} \times q_{tc}}_{\mathfrak{I}_t \times q_{tc}}} \underbrace{\begin{pmatrix} R_{t'_1 c'} E_{t'_1 c} \\ \vdots \\ R_{t'_\tau c'} E_{t'_\tau c} \end{pmatrix}}_{=: \widehat{V}_{tc} \in \mathbb{C}^{q_{tc} \times k_{tc}}}.$$

Die Matrix U_{tc} ist dabei erneut isometrisch. Sie entsteht durch Aneinanderfügen isometrischer Matrizen, die selbst jeweils Nichtnulleinträge in unterschiedlichen Zeilen haben, da die Kinder von t disjunkt sind.
Die Abbildung 5.2 zeigt am Beispiel von zwei Kindern, wie die Rekonstruktion von V_{tc} mit Hilfe von U_{tc} und $\widehat{V}_{tc}$ aussieht.

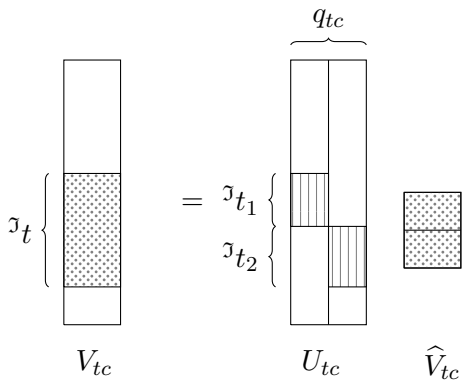


Abbildung 5.2: Rekonstruktion der Matrix V_{tc} der Clusterbasis im Nicht-Blattfall

Für $\widehat{V}_{tc}$ wird ebenfalls eine reduzierte QR-Zerlegung $\widehat{V}_{tc} = \widehat{Q}_{tc} R_{tc}$ bestimmt. Entsprechend existiert ein $p_{tc} = \min\{k_{tc}, q_{tc}\}$ mit $\widehat{Q}_{tc} \in \mathbb{C}^{q_{tc} \times p_{tc}}$ und $R_{tc} \in \mathbb{C}^{p_{tc} \times k_{tc}}$. Um anschließend

isometrische Transfermatrizen $E_{t'c}$ zu bestimmen, wird die Matrix $\widehat{Q}_{tc}$ als Ausgangsmatrix verwendet und entsprechend auf die Kinder aufgeteilt.
Auf die eben beschriebene Weise lässt sich ein rekursiver Algorithmus (5.1) zur Orthogonalisierung der richtungsabhängigen Clusterbasis gewinnen. Um diesen möglichst sparsam zu implementieren, werden die Matrizen der Clusterbasen direkt überschrieben und die für den Basiswechsel benötigten oberen Dreiecksmatrizen R_{tc} mit Hilfe einer richtungsabhängigen Menge von Matrizen R weitergereicht.

---

procedure ortho_dclusterbasis($\mathcal{T}_\mathcal{I}, \mathcal{R}, t, V, R$)
  Matrix $Q, \widehat{V}$, int $q, \hat{k}$
  if $\operatorname{kind}(t) = \emptyset$ **then**
    for $c \in \mathcal{R}_t$ **do**
      $\hat{k}_{tc} \leftarrow \min\left\{\#^\mathfrak{I} t, k_{tc}\right\}$, compute QR decomposition $V_{tc}|_{\star(t\times k_{tc})} = QR_{tc}$
      resize V_{tc}, $V_{tc}|_{\star(t\times\hat{k}_{tc})} \leftarrow Q$
    end for
  else
    for all $t' \in \operatorname{kind}(t)$ **do**
      ortho_dclusterbasis($\mathcal{T}_\mathcal{I}, \mathcal{R}, t', V, R$)
    end for
    for $c \in \mathcal{R}_t$ **do**
      $c' = \mathrm{r}_t(c) \in \mathcal{R}_{t'}$, $q \leftarrow \sum_{t'\in\operatorname{kind}(t)} \hat{k}_{t'c'}$, $\widehat{V} \in \mathbb{C}^{q\times k_{tc}}$
      $q \leftarrow 0$
      for $t' \in \operatorname{kind}(t)$ **do**
        $\widehat{V}|_{\star([q+1,q+\hat{k}_{t'c'}]\times k_{tc})} \leftarrow R_{t'c'}E_{t'c}$, $q \leftarrow q + \hat{k}_{t'c'}$
      end for
      $\hat{k}_{tc} \leftarrow \min\left\{k_{tc}, q\right\}$, compute QR decomposition $\widehat{V} = QR_{tc}$
      $q \leftarrow 0$
      for $t' \in \operatorname{kind}(t)$ **do**
        resize $E_{t'c}$, $E_{t'c} \leftarrow Q|_{\star([q+1,q+\hat{k}_{t'c'}]\times\hat{k}_{tc})}$, $q \leftarrow q + \hat{k}_{t'c'}$
      end for
    end for
  end if
end procedure

Algorithmus 5.1: Orthogonalisierung

Um den Aufwand der Orthogonalisierung abschätzen zu können, sind sowohl Schranken für den Aufwand der Matrix-Multiplikation als auch der QR-Zerlegung notwendig.

Bemerkung 29 (Aufwand Matrix-Multiplikation & QR-Zerlegung): *Der Aufwand der*

Multiplikation zweier Matrizen $A \in \mathbb{C}^{\mathcal{I}\times\mathcal{J}}$ und $B \in \mathbb{C}^{\mathcal{J}\times\mathcal{K}}$, wobei $\mathcal{I},\mathcal{J},\mathcal{K} \subset \mathbb{N}$ Indexmengen seien, beträgt

$$2(\#\mathcal{I})(\#\mathcal{J})(\#\mathcal{K}) \tag{5.1.3}$$

Operationen.
Es existiert eine Konstante $\mathcal{C}_{qr} \in \mathbb{N}$, so dass der Aufwand der QR-Zerlegung einer Matrix $A \in \mathbb{C}^{\mathcal{I}\times\mathcal{J}}$ durch

$$C_{qr}(\#\mathcal{I})(\#\mathcal{J}) \min\{\#\mathcal{I}, \#\mathcal{J}\} \tag{5.1.4}$$

beschränkt ist. Eine genaue Aufwandsbetrachtung der QR-Zerlegung findet sich zum Beispiel im Buch von Golub und Van Loan [26, S. 246 ff.].

Lemma 5.6 (Aufwand Orthogonalisierung Clusterbasis)
Sei $\{V_{tc}\}_{\substack{t\in\mathcal{T}_\mathcal{I}\\ c\in\mathcal{R}_t}}$ eine richtungsabhängige Clusterbasis zum Clusterbaum $\mathcal{T}_\mathcal{I}$ mit der Familie von Richtungsmengen $\mathcal{R}$ gegeben. Der Aufwand für die Orthogonalisierung aller Matrizen V_{tc} für $t \in \mathcal{T}_\mathcal{I}$, $c \in \mathcal{R}_t$ beträgt höchstens

$$k^2\mathcal{C}_{or}\left(\#\mathcal{I} + k\kappa^2\left(p_\mathcal{I}+1\right)\right), \tag{5.1.5}$$

dabei gilt

$$\mathcal{C}_{or} := \max\{(\mathcal{C}_{qr}+2)\mathcal{C}_{kk}, \mathcal{C}_{qr}\mathcal{C}_{bk}\}\max\{2\mathcal{C}_{bk}, \mathcal{C}_{tk}\}.$$

Beweis: Seien $t \in \mathcal{T}_\mathcal{I}$ und $c \in \mathcal{R}_t$ gegeben.
Betrachte zunächst den Fall, dass es sich bei $t \in \mathcal{T}_\mathcal{I}$ um ein Blattcluster handelt. Die Orthogonalisierung von $V_{tc} \in \mathbb{C}^{\mathcal{I}\times k_{tc}}_{\mathfrak{I}t\times k_{tc}}$ braucht nach (5.1.4) nicht mehr als

$$\mathcal{C}_{qr}(\#^\mathfrak{I}t)k_{tc}\min\left\{(\#^\mathfrak{I}t), k_{tc}\right\}$$

Operationen. Ohne Beschränkung der Allgemeinheit kann das Minimum gegen k_{tc} abgeschätzt werden. Somit lässt sich der Aufwand für die Orthogonalisierung durch $\mathcal{C}_{qr}(\#^\mathfrak{I}t)k_{tc}^2$ angeben. Für alle nötigen Richtungen zum Blattcluster t kann der Aufwand mit $\mathcal{R}_t^{eff}$ durch

$$\sum_{c\in\mathcal{R}_t^{eff}} \mathcal{C}_{qr}(\#^\mathfrak{I}t)k_{tc}^2 \leq \mathcal{C}_{qr}k^2(\#^\mathfrak{I}t)\sum_{c\in\mathcal{R}_t^{eff}} 1 \overset{(3.1.13)}{\leq} \mathcal{C}_{qr}k^3\mathcal{C}_{bk}\sum_{c\in\mathcal{R}_t^{eff}} 1$$

beschränkt werden.
Sei nun $t \in \mathcal{T}_\mathcal{I}\setminus\mathcal{L}_\mathcal{I}$. Zunächst muss für alle Kinder $t' \in \mathrm{kind}(t)$ in Richtung $c' = \mathrm{r}_t(c) \in \mathcal{R}_{t'}$ die Matrix $R_{t'c'}E_{t'c}$ mit $R_{t'c'} \in \mathbb{C}^{p_{t'c'}\times k_{t'c'}}$ für $p_{t'c'} = \min\{k_{t'c'}, q_{t'c'}\}$ und $E_{t'c} \in \mathbb{C}^{k_{t'c'}\times k_{tc}}$ berechnet werden. Dafür sind nach (5.1.3) $2k_{tc}p_{t'c'}k_{t'c'}$ Operationen nötig. Das Zusammenfügen der so entstehenden Teilmatrizen zu $\widehat{V}_{tc}$ kostet keine weiteren

Operationen. Sei $q_{tc} = \sum_{t' \in \operatorname{kind}(t)} p_{t'c'}$ die Anzahl der Zeilen von $\widehat{V}_{tc}$, dann ist der Aufwand der Orthogonalisierung der Matrix $\widehat{V}_{tc}$ durch $\mathcal{C}_{qr} q_{tc} k_{tc}^2$ beschränkt. Damit ergibt sich der Aufwand für eine Richtung mit

$$\begin{aligned}\mathcal{C}_{qr} q_{tc} k_{tc}^2 + \sum_{t' \in \operatorname{kind}(t)} 2 k_{tc} p_{t'c'} k_{t'c'} &= \sum_{t' \in \operatorname{kind}(t)} \mathcal{C}_{qr} p_{t'c'} k_{tc}^2 + \sum_{t' \in \operatorname{kind}(t)} 2 k_{tc} p_{t'c'} k_{t'c'} \\ &\leq (\mathcal{C}_{qr} + 2) k^2 \sum_{t' \in \operatorname{kind}(t)} p_{t'c'}.\end{aligned}$$

Da $p_{t'c'}$ sich über die Minimumsbildung durch $k_{t'c'}$ abschätzen lässt, ergibt sich mit der Kinderkonstante 3.9

$$(\mathcal{C}_{qr} + 2) k^2 \sum_{t' \in \operatorname{kind}(t)} p_{t'c'} \leq (\mathcal{C}_{qr} + 2) k^3 \mathcal{C}_{kk}$$

und für alle Richtungen $c \in \mathcal{R}_t^{eff}$ wird dies zu

$$\sum_{c \in \mathcal{R}_t^{eff}} (\mathcal{C}_{qr} + 2) k^3 \mathcal{C}_{kk}.$$

Sei $\mathcal{C}_o := \max\{(\mathcal{C}_{qr} + 2)\mathcal{C}_{kk}, \mathcal{C}_{qr}\mathcal{C}_{bk}\}$, dann folgt für die gesamte Clusterbasis

$$\begin{aligned}\sum_{t \in \mathcal{L}_\mathcal{I}} \mathcal{C}_{qr} k^3 \mathcal{C}_{bk} \sum_{c \in \mathcal{R}_t^{eff}} 1 + \sum_{t \in \mathcal{T}_\mathcal{I} \setminus \mathcal{L}_\mathcal{I}} \sum_{c \in \mathcal{R}_t^{eff}} (\mathcal{C}_{qr} + 2) k^3 \mathcal{C}_{kk} \\ \leq \mathcal{C}_o k^3 \sum_{t \in \mathcal{T}_\mathcal{I}} \sum_{c \in \mathcal{R}_t^{eff}} 1.\end{aligned}$$

Dies kann mit Lemma 3.14 und Korollar 3.17 weiter abgeschätzt werden

$$\begin{aligned}\mathcal{C}_o k^3 \sum_{t \in \mathcal{T}_\mathcal{I}} \sum_{c \in \mathcal{R}_t^{eff}} 1 &\leq \mathcal{C}_o k^3 \left(\#\mathcal{T}_\mathcal{I} + \kappa^2 (p_\mathcal{I} + 1) \mathcal{C}_{tk} \right) \\ &\leq \mathcal{C}_o k^2 \left(2\mathcal{C}_{bk} \#\mathcal{I} + k\kappa^2 (p_\mathcal{I} + 1) \mathcal{C}_{tk} \right).\end{aligned}$$

□

Mit Bemerkung 12 ergibt sich ein Aufwand von $\mathcal{O}(k^2 \#\mathcal{I} + k^3 \kappa^2 \log_2(\#\mathcal{I}))$.

Selbst wenn durch das Orthogonalisieren nicht in allen Fällen Speicher eingespart werden kann, bieten isometrische Clusterbasen einen anderen Vorteil. Mit der orthogonalen Projektion ist es möglich, die Bestapproximation einer Matrix unter Verwendung dieser Clusterbasis zu erhalten. Dies ist die Grundidee, auf der die direkte Kompression und die Rekompression von $\mathcal{RH}^2$-Matrizen beruhen.

5.2 Direkte Kompression

Bei der direkten Kompression wird eine $\mathcal{RH}^2$-Matrix-Approximation einer gegebenen Matrix A generiert, ohne auf Interpolation als Hilfsmittel zurückzugreifen. Dafür sind der richtungsabhängige Clusterbaum $\mathcal{T}_{\mathcal{I}}$, die Familie von Richtungsmengen $\mathcal{R}$ sowie der richtungsabhängige Blockbaum $\mathcal{T}_{\mathcal{I}\times\mathcal{I}}$ notwendig. Der Algorithmus stammt von Börm [6] und ist eine Abwandlung des Algorithmus zur direkten Kompression von $\mathcal{H}^2$-Matrizen.
Die orthogonale Projektion spielt eine Schlüsselrolle bei diesem Kompressionsansatz, denn wenn die richtungsabhängigen Clusterbasen isometrisch sind, handelt es sich bei $V_{tc}V_{tc}^*$ beziehungsweise $W_{sc}W_{sc}^*$ um orthogonale Projektionen. Setze für einen zulässigen Block $b = (t, s, c)$ die Kopplungsmatrix mit $S_b := V_{tc}^* A|_{t\times s} W_{sc}$, dann ist durch

$$V_{tc}V_{tc}^* A|_{t\times s} W_{sc}W_{sc}^* = V_{tc}S_b W_{sc}^* \tag{5.2.1}$$

schon eine Approximation der Matrix $A|_{t\times s} \in \mathbb{C}^{\mathcal{I}\times\mathcal{I}}_{\mathfrak{I}_t\times\mathfrak{I}_s}$ gegeben. Der Approximationsfehler lässt sich auf die Fehler der einzelnen Projektionen zurückführen und darüber kontrollieren. Für die Spektralnorm ebenso wie für die Frobeniusnorm gilt

$$\begin{aligned}&\|A|_{t\times s} - V_{tc}V_{tc}^* A|_{t\times s} W_{sc}W_{sc}^*\|^2\\ &\qquad = \|A|_{t\times s} - V_{tc}V_{tc}^* A|_{t\times s} + V_{tc}V_{tc}^*(A|_{t\times s} - A|_{t\times s} W_{sc}W_{sc}^*)\|^2.\end{aligned}$$

Für die Frobeniusnorm folgt direkt

$$\begin{aligned}&\|A|_{t\times s} - V_{tc}V_{tc}^* A|_{t\times s} + V_{tc}V_{tc}^*(A|_{t\times s} - A|_{t\times s} W_{sc}W_{sc}^*)\|_F^2\\ &\qquad = \|A|_{t\times s} - V_{tc}V_{tc}^* A|_{t\times s}\|_F^2 + \|V_{tc}V_{tc}^*(A|_{t\times s} - A|_{t\times s} W_{sc}W_{sc}^*)\|_F^2\\ &\qquad + \langle A|_{t\times s} - V_{tc}V_{tc}^* A|_{t\times s}, V_{tc}V_{tc}^*(A|_{t\times s} - A|_{t\times s} W_{sc}W_{sc}^*)\rangle_F\\ &\qquad + \langle V_{tc}V_{tc}^*(A|_{t\times s} - A|_{t\times s} W_{sc}W_{sc}^*), A|_{t\times s} - V_{tc}V_{tc}^* A|_{t\times s}\rangle_F,\end{aligned}$$

wobei die Skalarprodukte aufgrund der Eigenschaft der orthogonalen Projektion wegfallen, denn es gilt

$$\langle A|_{t\times s} - V_{tc}V_{tc}^* A|_{t\times s}, V_{tc}V_{tc}^* U\rangle_F = 0 \qquad \text{für alle } U \in \mathbb{C}^{\mathcal{I}\times\mathcal{I}}_{\mathfrak{I}_t\times\mathfrak{I}_s}.$$

Da orthogonale Projektionen die Norm nicht vergrößern, folgt schließlich

$$\begin{aligned}&\|A|_{t\times s} - V_{tc}V_{tc}^* A|_{t\times s} W_{sc}W_{sc}^*\|_F^2\\ &\qquad = \|A|_{t\times s} - V_{tc}V_{tc}^* A|_{t\times s}\|_F^2 + \|V_{tc}V_{tc}^*(A|_{t\times s} - A|_{t\times s} W_{sc}W_{sc}^*)\|_F^2\\ &\qquad \leq \|A|_{t\times s} - V_{tc}V_{tc}^* A|_{t\times s}\|_F^2 + \|A|_{t\times s} - A|_{t\times s} W_{sc}W_{sc}^*\|_F^2.\end{aligned}$$

Im Fall der Spektralnorm läuft die Argumentation analog, ist aufgrund der Maximumsbildung bei der Betrachtung der Skalarprodukte nur umfangreicher aufzuschreiben

$$\begin{aligned}
&\|A|_{t\times s} - V_{tc}V_{tc}^*A|_{t\times s} + V_{tc}V_{tc}^*(A|_{t\times s} - A|_{t\times s}W_{sc}W_{sc}^*)\|_2^2 \\
&\quad\leq \|A|_{t\times s} - V_{tc}V_{tc}^*A|_{t\times s}\|_2^2 + \|V_{tc}V_{tc}^*(A|_{t\times s} - A|_{t\times s}W_{sc}W_{sc}^*)\|_2^2 \\
&\quad+ \max_{x\in\mathbb{C}^{\mathcal{I}},\|x\|_2=1} \langle (A|_{t\times s} - V_{tc}V_{tc}^*A|_{t\times s})x, V_{tc}V_{tc}^*(A|_{t\times s} - A|_{t\times s}W_{sc}W_{sc}^*)x\rangle_2 \\
&\quad+ \max_{x\in\mathbb{C}^{\mathcal{I}},\|x\|_2=1} \langle V_{tc}V_{tc}^*(A|_{t\times s} - A|_{t\times s}W_{sc}W_{sc}^*)x, (A|_{t\times s} - V_{tc}V_{tc}^*A|_{t\times s})x\rangle_2.
\end{aligned}$$

Auch hier fallen die Skalarprodukte aufgrund der orthogonalen Projektionen weg, da orthogonale Projektionen die Spektralnorm nicht vergrößern, folgt erneut

$$\begin{aligned}
&\|A|_{t\times s} - V_{tc}V_{tc}^*A|_{t\times s} + V_{tc}V_{tc}^*(A|_{t\times s} - A|_{t\times s}W_{sc}W_{sc}^*)\|_2^2 \\
&\quad\leq \|A|_{t\times s} - V_{tc}V_{tc}^*A|_{t\times s}\|_2^2 + \|V_{tc}V_{tc}^*(A|_{t\times s} - A|_{t\times s}W_{sc}W_{sc}^*)\|_2^2 \\
&\quad\leq \|A|_{t\times s} - V_{tc}V_{tc}^*A|_{t\times s}\|_2^2 + \|A|_{t\times s} - A|_{t\times s}W_{sc}W_{sc}^*\|_2^2.
\end{aligned}$$

Damit kann der Fehler für die Spektral- und Frobeniusnorm durch die Summe der Fehler der einzelnen Projektionen abgeschätzt werden

$$\begin{aligned}
&\|A|_{t\times s} - V_{tc}V_{tc}^*A|_{t\times s}W_{sc}W_{sc}^*\|^2 \\
&\quad\leq \|A|_{t\times s} - V_{tc}V_{tc}^*A|_{t\times s}\|^2 + \|A|_{t\times s}^* - W_{sc}W_{sc}^*A|_{t\times s}^*\|^2.
\end{aligned}$$

Folglich gilt es, den Fehler der Projektion unter Kontrolle zu bringen sowie Clusterbasen zu finden, die einen möglichst kleinen Rang haben. Die Verwendung einer Singulärwertzerlegung liefert beide gewünschten Eigenschaften und ermöglicht es sogar, in Bezug auf Spektral- und Frobeniusnorm bei festgelegter Genauigkeit den optimalen Rang zu wählen.

Definition und Lemma 5.7 (Singulärwertzerlegung)
Zu jeder Matrix $A \in \mathbb{C}^{M\times N}\setminus\{0\}$ existiert ein $p \leq \min\{\#M, \#N\}$ mit $p > 0$ und geordneten positiven Singulärwerten $\sigma_1 \geq \cdots \geq \sigma_p$. *Weiter existieren zwei isometrische Matrizen $O \in \mathbb{C}^{N\times p}, U \in \mathbb{C}^{M\times p}$, so dass*

$$A = U\Sigma O^* \qquad \textit{mit} \qquad \mathbb{R}^{p\times p} \ni \Sigma := \operatorname{diag}(\sigma_1, \ldots, \sigma_p)$$

gilt. Die Spaltenvektoren von U werden als linke und die Spaltenvektoren von O als rechte Singulärvektoren *bezeichnet.*

Für einen Existenzbeweis im reellen Fall, der sich auf den komplexen übertragen lässt, sei der Leser auf [26, §2.4] verwiesen. Der Aufwand der Singulärwertzerlegung einer quadratischen Matrix $\overline{A} \in \mathbb{C}^{M\times M}$ liegt in $\mathcal{O}((\#M)^3)$ (Vergleich siehe [26, S. 293]). Entsprechend existiert eine Konstante $\mathcal{C}_{svd} \in \mathbb{R}_{>0}$ derart, dass der Aufwand einer Singulärwertzerlegung[ii]

[ii] Bezieht sich auf eine Bestimmung bis auf Maschinengenauigkeit, dabei geht der Singulärwertzerlegung meist eine QR-Zerlegung voran, um die Matrix für die Singulärwertzerlegung auf quadratische Gestalt zu bringen.

einer Matrix $A \in \mathbb{C}^{M\times N}$ zu einer festen Genauigkeit $\epsilon \in \mathbb{R}_{>0}$ durch

$$\mathcal{C}_{svd}\,(\#M)(\#N)\min\{\#M,\#N\} \tag{5.2.2}$$

beschränkt ist.
Eine Singulärwertzerlegung kann verwendet werden, um den Rang zu reduzieren. Angenommen zu einer Matrix $A \in \mathbb{C}^{M\times N}$ liegt eine Singulärwertzerlegung vor, es existieren also zwei isometrische Matrizen $U \in \mathbb{C}^{M\times q}$ und $O \in \mathbb{C}^{N\times q}$ sowie $q \leq \min\{\#M,\#N\}$ mit

$$A = U\,\mathrm{diag}(\sigma_1,\ldots,\sigma_q)O^*,$$

wobei $\sigma_1 \geq \cdots \geq \sigma_q$ gilt. Der Gedanke ist nun, σ_q gleich null zu setzen und zu betrachten, ob mit der so entstehenden Matrix $\widetilde{A}$ die vorgegebene Fehlerschranke noch eingehalten wird. Ist dies der Fall, wird das Vorgehen mit σ_{q-1} wiederholt. Solange die Fehlerschranke eingehalten wird, kann dieses Vorgehen fortgeführt werden. Folglich wird das minimale $\ell \in \underline{q}_|$ gesucht, bei dem der Fehler zwischen A und $\widetilde{A}$ noch tolerierbar ist.
Natürlich ist es nicht praktikabel, die Singulärwerte nach und nach zu löschen und dann jeweils den Fehler zu überprüfen. Zum Glück ist dies jedoch auch nicht notwendig.
Sei $\ell \in \underline{q}_|$ und $\hat{U}$ bestehe aus den ersten ℓ Spalten von U, also $\hat{U} = U|_{M\times[1,\ell]}$. Dann gilt

$$\widetilde{A} := \hat{U}\hat{U}^*A = U\,\mathrm{diag}(\sigma_1,\ldots,\sigma_\ell,0,\ldots,0)O^*,$$

denn es gelten die folgende Umformungen

$$\begin{aligned}\hat{U}^*A &= (U^*A)|_{[1,\ell]\times N} = (U^*U\,\mathrm{diag}(\sigma_1,\ldots,\sigma_q)O^*)|_{[1,\ell]\times N}\\ &= (\mathrm{diag}(\sigma_1,\ldots,\sigma_q)O^*)\,|_{[1,\ell]\times N} = (\mathrm{diag}(\sigma_1,\ldots,\sigma_\ell,0,\ldots,0)O^*)\,|_{[1,\ell]\times N}.\end{aligned}$$

Damit folgt

$$\begin{aligned}\|A-\hat{U}\hat{U}^*A\| &= \|U\,\mathrm{diag}(\sigma_1,\ldots,\sigma_q)O^* - U\,\mathrm{diag}(\sigma_1,\ldots,\sigma_\ell,0,\ldots,0)O^*\|\\ &= \|\,\mathrm{diag}(\sigma_1,\ldots,\sigma_q) - \mathrm{diag}(\sigma_1,\ldots,\sigma_\ell,0,\ldots,0)\|\\ &= \|\,\mathrm{diag}(0,\ldots,0,\sigma_{\ell+1},\ldots,\sigma_q)\|\end{aligned}$$

sowohl für die Spektral- als auch für die Frobeniusnorm. Insgesamt ergibt sich so ein Fehler von

$$\|A-\hat{U}\hat{U}^*A\| = \begin{cases}\sigma_{\ell+1} & \text{für } \|\cdot\|_2,\\ \left(\sum\limits_{i=\ell+1}^{q}\sigma_i^2\right)^{\frac{1}{2}} & \text{für } \|\cdot\|_F.\end{cases} \tag{5.2.3}$$

Entsprechend reicht es aus, die Genauigkeit mit Hilfe der Singulärwerte auf oben beschriebene Weise zu überprüfen. Zudem genügt es in der Praxis, die Berechnung der Singulärvektoren auf die linken zu beschränken, da nur diese für die orthogonale Projektion notwendig

sind und für die Berechnung der Spaltenclusterbasis die adjungierte Matrix A^* betrachtet werden kann. Die gekürzte Matrix der linken Singulärvektoren liefert die Grundlage für die richtungsabhängige Clusterbasis. Da der Aufwand des Kürzens linear von q abhängt und nur die linken Singulärvektoren berechnet werden müssen, kann auch der Aufwand, eine gekürzte Singulärwertzerlegung zu bestimmen, mit einem $\mathcal{C}_{svd} \in \mathbb{R}$ wie in (5.2.2) beschränkt werden.
Da die Clusterbasen eine Schachtelungseigenschaft zu erfüllen haben, wird die Aufgabe erschwert und es stellt sich die Frage, wie die Schachtelungseigenschaft sichergestellt werden kann. Um diese Frage zu beantworten, reicht es vollkommen aus, sich auf die Zeilenclusterbasis zu konzentrieren, da die Spaltenclusterbasis analog durch Betrachtung von A^* gewonnen werden kann.

Seien ein $t \in \mathcal{T}_\mathcal{I}$, eine Richtung $c \in \mathcal{R}_t$ sowie eine Genauigkeit $\epsilon \in \mathbb{R}_{>0}$ gegeben. Aufgrund der Schachtelungseigenschaft der Clusterbasen müssen für die Bestimmung der Matrix V_{tc} auch Teilmatrizen von A betrachtet werden, die zu zulässigen Blöcken gehören, die mit Vorfahren von t gebildet werden. Demnach reicht die Betrachtung der direkten Vorfahrenrichtung allein nicht immer aus, weshalb für die Richtungen ein Analogon zur Menge $\mathrm{vor}\,(\cdot)$ definiert wird.

Definition 5.8 (Vorfahrenrichtungen)
Zu einem Cluster $t \in \mathcal{T}_\mathcal{I}$ und einer Richtung $c \in \mathcal{R}_t$ sind die Vorfahren der Richtung c zusammengefasst in der Menge

$$\mathrm{vor}_t(c) := \begin{cases} \{c\} & \text{für } \mathrm{vor}(t) = \{t\}, \\ \{c\} \cup \bigcup\limits_{c^+ \in \mathrm{r}_{t^+}^{-1}(c)} \mathrm{vor}_{t^+}(c^+) & \text{sonst mit } t \in \mathrm{kind}(t^+). \end{cases}$$

Die Vorfahrenrichtungen für einen festen Cluster-Vorfahren $t^+ \in \mathrm{vor}(t)$ sind durch die Menge

$$\mathrm{vor}_t^{t^+}(c) := \{\hat{c} \in \mathrm{vor}_t(c) \mid \hat{c} \in \mathcal{R}_{t^+}\}$$

gegeben, falls $t \in \mathrm{kind}(t^+)$ gilt, entspricht $\mathrm{vor}_t^{t^+}(c) = \mathrm{r}_{t^+}^{-1}(c)$.

Fasse alle Spaltencluster von zulässigen Teilmatrizen von A, die eine Verbindung zu dem betrachteten t haben, in

$$S^{tc} := \left\{ s \in \mathcal{T}_\mathcal{I} \mid \exists t^+ \in \mathrm{vor}(t), c_b \in \mathrm{vor}_t^{t^+}(c) \text{ mit } (t^+, s, c_b) \in \mathcal{L}^+_{\mathcal{I}\times\mathcal{I}} \right\} \tag{5.2.4}$$

sowie die dazugehörigen Indizes in

$$\mathcal{S}_{tc} := \bigcup_{s \in S^{tc}} {}^\mathfrak{I}s \tag{5.2.5}$$

zusammen und die Mächtigkeit dieser Menge in $s_{tc} := \#\mathcal{S}_{tc}$. Bezeichne den damit definierten Teil der Matrix A mit $A^{tc} \in \mathbb{C}^{\mathcal{I}\times\mathcal{S}_{tc}}_{\mathfrak{I}t\times\mathcal{S}_{tc}}$ mit

$$A^{tc}|_{\star(t\times\mathcal{S}_{tc})} := A|_{\star(t\times\mathcal{S}_{tc})}.$$

Im Fall, dass es sich bei t um ein Blattcluster handelt, kann direkt die eben definierte Matrix A^{tc} verwendet und die Matrix ihrer linken Singulärvektoren auf die notwendige Größe gekürzt werden. Sei durch $q \in \underline{\min\left\{\#^{\mathfrak{I}}t, s_{tc}\right\}_{\rfloor}}$ und

$$A^{tc}|_{\star(t\times\mathcal{S}_{tc})} = U\,\mathrm{diag}(\sigma_1,\ldots,\sigma_q)O^*$$

die Singulärwertzerlegung mit isometrischen Matrizen $U \in \mathbb{C}^{\mathfrak{I}t\times q}$, $O \in \mathbb{C}^{s_{tc}\times q}$ gegeben. Bezeichne mit k_{tc} den minimalen Rang, bei dem der Fehler noch im Bereich der Genauigkeit ϵ liegt, dann ist die gesuchte isometrische Matrix $V_{tc} \in \mathbb{C}^{\mathcal{I}\times k_{tc}}_{\mathfrak{I}t\times k_{tc}}$ gegeben durch $V_{tc}|_{\star(t\times k_{tc})} = U|_{t\times k_{tc}}$.
Handelt es sich bei t nicht um ein Blattcluster, gilt es, die geschachtelte Struktur zu bedenken. Nehme an, dass für die Kinder von t zur Richtung $c' = \mathrm{r}_t(c)$ die Matrizen der Clusterbasis bereits berechnet wurden. Sei $\tau = \#\,\mathrm{kind}(t)$. Gesucht sind dann Transfermatrizen $E_{t_1c},\ldots,E_{t_\tau c}$, so dass $V_{tc} = \sum_{i=1}^{\tau} V_{t_ic'}E_{t_ic}$ gilt und die Bedingung

$$\|A^{tc} - V_{tc}V_{tc}^*A^{tc}\| \leq \epsilon$$

erfüllt ist. Aufteilen der Matrix A^{tc} in die Anteile der Kinder von t und Substituieren der ersten Matrix V_{tc} führt zu

$$A^{tc} - V_{tc}V_{tc}^*A^{tc} = \sum_{i=1}^{\tau} A^{tc}|_{t_i\times\mathcal{S}_{tc}} - \sum_{i=1}^{\tau} V_{t_ic'}E_{t_ic}V_{tc}^*A^{tc}$$

und schließlich ergibt sich

$$\|A^{tc} - V_{tc}V_{tc}^*A^{tc}\|^2 \leq \sum_{i=1}^{\tau} \left\|A^{tc}|_{t_i\times\mathcal{S}_{tc}} - V_{t_ic'}E_{t_ic}V_{tc}^*A^{tc}\right\|^2.$$

Füge für alle $i \in \underline{\tau_{\rfloor}}$ eine nahrhafte Null hinzu

$$\|A^{tc}|_{t_i\times\mathcal{S}_{tc}} - V_{t_ic'}V_{t_ic'}^*A^{tc}|_{t_i\times\mathcal{S}_{tc}} + V_{t_ic'}\left(V_{t_ic'}^*A^{tc}|_{t_i\times\mathcal{S}_{tc}} - E_{t_ic}V_{tc}^*A^{tc}\right)\|^2,$$

dann entfallen beim Auseinanderziehen der Normen die gemischten Terme, was zu einer oberen Schranke von

$$\|A^{tc}|_{t_i\times\mathcal{S}_{tc}} - V_{t_ic'}V_{t_ic'}^*A^{tc}|_{t_i\times\mathcal{S}_{tc}}\|^2 + \|V_{t_ic'}\left(V_{t_ic'}^*A^{tc}|_{t_i\times\mathcal{S}_{tc}} - E_{t_ic}V_{tc}^*A^{tc}\right)\|^2$$

führt. Dass die gemischten Terme entfallen, ist leicht am Beispiel des Frobeniusskalarprodukts einzusehen

$$\begin{aligned}&\langle A^{tc}|_{t_i\times\mathcal{S}_{tc}} - V_{t_ic'}V_{t_ic'}^*A^{tc}|_{t_i\times\mathcal{S}_{tc}}, V_{t_ic'}\left(V_{t_ic'}^*A^{tc}|_{t_i\times\mathcal{S}_{tc}} - E_{t_ic}V_{tc}^*A^{tc}\right)\rangle_F\\ &= \langle 0, V_{t_ic'}^*A^{tc}|_{t_i\times\mathcal{S}_{tc}} - E_{t_ic}V_{tc}^*A^{tc}\rangle_F.\end{aligned}$$

Über die erste Norm ist bekannt, dass sie für alle $i \in \underline{\tau}$ die ϵ-Schranke einhält. Bleibt nur die zweite Norm zu untersuchen. Hier kann die ausgeklammerte Matrix $V_{t_ic'}$ weggelassen werden, da sie für alle i eine isometrische Matrix ist. Zudem können die Normen für alle i wieder zusammengefasst werden, da hier Blockmatrizen, welche mit nullen aufgefüllt sind, betrachtet werden. Bei diesem Schritt wird auch gleich die verbliebene Matrix V_{tc}^* durch die Darstellung mit ihren Kinder ersetzt

$$\Bigg\| \underbrace{\begin{pmatrix} V_{t_1c'}^*A^{tc}|_{t_1\times\mathcal{S}_{tc}} \\ \vdots \\ V_{t_\tau c'}^*A^{tc}|_{t_\tau\times\mathcal{S}_{tc}} \end{pmatrix}}_{=:\widehat{A}^{tc}} - \underbrace{\begin{pmatrix} E_{t_1c} \\ \vdots \\ E_{t_\tau c} \end{pmatrix}}_{=:\widehat{V}_{tc}} \underbrace{\begin{pmatrix} E_{t_1c} \\ \vdots \\ E_{t_\tau c} \end{pmatrix}^*}_{\widehat{V}_{tc}^*} \underbrace{\begin{pmatrix} V_{t_1c'}^*A^{tc}|_{t_1\times\mathcal{S}_{tc}} \\ \vdots \\ V_{t_\tau c'}^*A^{tc}|_{t_\tau\times\mathcal{S}_{tc}} \end{pmatrix}}_{\widehat{A}^{tc}} \Bigg\|^2.$$

Auf diese Weise kann die gleiche Struktur wie beim Ausgangsproblem erhalten werden. Demnach gilt es nun, $\widehat{V}_{tc}$ mit Hilfe der Singulärwertzerlegung von $\widehat{A}^{tc}$ so zu bestimmen, dass die Schranke von ϵ eingehalten und der Rang minimiert wird.
Im Hinblick auf einen effizienten Algorithmus ist es sinnvoll, nachdem die neue Matrix der Clusterbasis bestimmt wurde, $R_{tc} = V_{tc}^*A^{tc}$ vorzuberechnen und zu speichern. Im nächsten Schritt der Rekursion kann $\widehat{A}^{tc}$ so einfach über Zusammenkopieren und Einschränken erstellt werden.

Um im Endergebnis eine gewisse Genauigkeit für die Kompression gewährleisten zu können, ist es hilfreich, die Fehlerschranke für jeden Cluster einzeln festzulegen, also für den Cluster $t \in \mathcal{T}_\mathcal{I}$ eine Schranke von $\epsilon_t > 0$ zu fordern

$$\|A^{tc} - V_{tc}V_{tc}^*A^{tc}\|^2 \leq \epsilon_t^2.$$

Der Gesamtfehler kann dann, wie in [6, Thm. 15] erwähnt, mit dem folgenden Lemma beschränkt und auch gezielt kontrolliert werden.

Lemma 5.9 Fehler
Existiere eine Familie $\{\epsilon_t\}_{t\in\mathcal{T}_\mathcal{I}}$, so dass

$$\begin{aligned}\|A^{tc} - V_{tc}V_{tc}^*A^{tc}\|^2 &\leq \epsilon_t^2 && \text{für alle } t \in \mathcal{L}_\mathcal{I}, c \in \mathcal{R}_t\\ \|\widehat{A}^{tc} - \widehat{V}_{tc}\widehat{V}_{tc}^*\widehat{A}^{tc}\|^2 &\leq \epsilon_t^2 && \text{für alle } t \in \mathcal{T}_\mathcal{I} \setminus \mathcal{L}_\mathcal{I}, c \in \mathcal{R}_t\end{aligned}$$

gilt, kann der Fehler der orthogonalen Projektion für alle $t \in \mathcal{T}_\mathcal{I}$ und $c \in \mathcal{R}_t$ durch

$$\|A^{tc} - V_{tc}V_{tc}^* A^{tc}\|^2 \leq \sum_{t' \in \mathrm{nac}\,(t)} \epsilon_{t'}^2$$

beschränkt werden.

Beweis: Per abschnittsweiser Induktion.
I.A. Im ersten Schritt sei $t \in \mathcal{T}_\mathcal{I}$ mit $\mathrm{stufe}(t) = p_\mathcal{I}$ gegeben, dann ist $t \in \mathcal{L}_\mathcal{I}$ und es gilt nach Voraussetzung für alle $c \in \mathcal{R}_t$

$$\|A^{tc} - V_{tc}V_{tc}^* A^{tc}\|^2 \leq \epsilon_t^2.$$

I.V. Sei $n \in \underline{p_\mathcal{I}|}$ gegeben, so dass die Behauptung für alle $t' \in \mathcal{T}_\mathcal{I}$ mit $\mathrm{stufe}(t') \geq n$ gilt.
I.S. Sei $t \in \mathcal{T}_\mathcal{I}$ mit $\mathrm{stufe}(t) = n-1$ gegeben. Falls $t \in \mathcal{L}_\mathcal{I}$ gilt, folgt die Behauptung sofort aus den Voraussetzungen. Ansonsten ist für alle Kinder $t' \in \mathrm{kind}\,(t)$ die Behauptung nach Induktionsvoraussetzung erfüllt, denn es gilt $\mathrm{stufe}(t') = n$. Damit folgt für den Fehler nach obiger Betrachtung mit $c' = \mathrm{r}_t(c)$

$$\|A^{tc} - V_{tc}V_{tc}^* A^{tc}\|^2 \leq \sum_{t' \in \mathrm{kind}(t)} \|A^{tc}|_{t' \times \mathcal{S}_{tc}} - V_{t'c'}V_{t'c'}^* A^{tc}|_{t' \times \mathcal{S}_{tc}}\|^2 + \|\widehat{A}^{tc} - \widehat{V}_{tc}\widehat{V}_{tc}^* \widehat{A}^{tc}\|^2.$$

Da $A^{tc}|_{t' \times \mathcal{S}_{tc}}$ eine Teilmatrix von $A^{t'c'}$ ist und alle Kinder $t' \in \mathrm{kind}(t)$ die Induktionsvoraussetzung erfüllen, wird der Fehler der verbleibenden Norm nach Voraussetzung kontrolliert. Es ergibt sich

$$\|A^{tc} - V_{tc}V_{tc}^* A^{tc}\|^2 \leq \sum_{t' \in \mathrm{nac}\,(t)} \epsilon_{t'}^2.$$

□

Eine geschickte Wahl von ϵ_t kann in Lemma 5.9 zu einer geometrischen Summe auf der rechten Seite und damit zu einem unabhängig von t beschränkten Fehler führen (vgl. [6, Bem. 16]). Es besteht auch die Möglichkeit, über Gewichtungsfaktoren eine verfeinerte Fehlerkontrolle durchzuführen, zum Beispiel um blockrelative Fehler zu gewährleisten.
Da nun sichergestellt ist, dass der Fehler kontrolliert werden kann, ist es Zeit, den konkreten Algorithmus für die direkte Kompression aufzustellen.

Beim Abschätzen des Aufwands des Algorithmus für die richtungsabhängigen Clusterbasen zeigt sich, dass aufgrund des häufigen Einsatzes der Singulärwertzerlegung mit keinem besonders schnellen Verfahren gerechnet werden kann. Der Aufwand für die Verarbeitung von $(\#\mathcal{I})^2$ Daten befindet sich mit $\mathcal{O}(k(\#\mathcal{I})^2)$ im zu erwartenden Rahmen.

Lemma 5.10 (Aufwand Berechnung Clusterbasis)
Der Aufwand zur Bestimmung einer richtungsabhängigen Clusterbasis zu einer gegebenen

procedure compress_bases($\mathcal{T}_\mathcal{I}, \mathcal{R}, t, A, \epsilon, V, R$)
  Matrix $A^{tc}, \widehat{A}^{tc}, \widehat{V}_{tc}, U, \Sigma, O,$ int k
  if $\operatorname{kind}(t) = \emptyset$ **then**
    for $c \in \mathcal{R}_t$ **do**
      $A^{tc} \leftarrow A|_{\star(t\times\mathcal{S}_{tc})}$, compute truncated SVD of $A^{tc} = U\Sigma O^*$ with rank k_{tc}, $V_{tc}|_{\star(t\times k_{tc})} \leftarrow U|_{\star(t\times k_{tc})}$, $R_{tc} \leftarrow (V_{tc}|_{\star(t\times k_{tc})})^* A^{tc}$
    end for
  else
    for $t' \in \operatorname{kind}(t)$ **do**
      compress_bases($\mathcal{T}_\mathcal{I}, \mathcal{R}, t', A, \epsilon, V, R$)
    end for
    for $c \in \mathcal{R}_t$ **do**
      $c' = \mathrm{r}_t(c) \in \mathcal{R}_{t'}$, $k \leftarrow \sum_{t'\in\operatorname{kind}(t)} k_{t'c'}$, $\widehat{A}^{tc} \in \mathbb{C}^{k\times\mathcal{S}_{tc}}$
      $k \leftarrow 0$
      for $t' \in \operatorname{kind}(t)$ **do**
        $\widehat{A}^{tc}|_{\star([k+1,k+k_{t'c'}]\times\mathcal{S}_{tc})} \leftarrow R_{t'c'}|_{\star(k_{t'c'}\times\mathcal{S}_{tc})}$, $k \leftarrow k + k_{t'c'}$
      end for
      Compute truncated SVD of $\widehat{A}^{tc} = U\Sigma O^*$ with rank k_{tc}
      $\widehat{V}_{tc} \leftarrow U|_{\star(k\times k_{tc})}$, $R_{tc} \leftarrow \widehat{V}_{tc}^* \widehat{A}^{tc}$
      Copy transfer matrices $E_{t'c}$ for all $t' \in \operatorname{kind}(t)$ out of $\widehat{V}_{tc}$
    end for
  end if
end procedure

Algorithmus 5.2: Berechnung der komprimierten Clusterbasis

Matrix $A \in \mathbb{C}^{\mathcal{I}\times\mathcal{I}}$ mit festgelegtem richtungsabhängigen Clusterbaum $\mathcal{T}_\mathcal{I}$ sowie Blockbaum $\mathcal{T}_{\mathcal{I}\times\mathcal{I}}$ beträgt weniger als

$$\mathcal{C}_{ccb} k (\#\mathcal{I})^2$$

Operationen mit

$$\mathcal{C}_{ccb} = 2\mathcal{C}_{bk} \max\left\{\mathcal{C}_{svd}\mathcal{C}_{kk}^2 + 2\mathcal{C}_{kk}, \mathcal{C}_{svd}\mathcal{C}_{bk}^2 + 2\mathcal{C}_{bk}\right\}.$$

Beweis: Betrachte zunächst den Fall eines Blattclusters $t \in \mathcal{T}_\mathcal{I}$ für ein $c \in \mathcal{R}_t$. Hier kann direkt mit der Singulärwertzerlegung gearbeitet werden, so dass sich mit der Komplexität der SVD (5.2.2) sowie der Annahme zur Auflösung der Blätter (3.1.13) ein Aufwand von

$$\mathcal{C}_{svd}(\#\mathfrak{I}t)s_{tc} \min\left\{(\#\mathfrak{I}t), s_{tc}\right\} \leq \mathcal{C}_{svd}(\#\mathfrak{I}t)^2 s_{tc} \leq \mathcal{C}_{svd} k^2 \mathcal{C}_{bk}^2 s_{tc}$$

ergibt. Zusätzlich ist noch die vorbereitende Multiplikation zur Bestimmung von R_{tc} abzuschätzen, für die ein Aufwand von

$$2k_{tc}(\#^{\mathfrak{I}}t)s_{tc} \leq 2k^2\mathcal{C}_{bk}s_{tc}$$

erforderlich ist. Insgesamt ergibt sich damit für den Cluster t ein Aufwand von

$$\sum_{c\in\mathcal{R}_t}(\mathcal{C}_{svd}\mathcal{C}_{bk}^2+2\mathcal{C}_{bk})k^2 s_{tc} = (\mathcal{C}_{svd}\mathcal{C}_{bk}^2+2\mathcal{C}_{bk})k^2\#\bigcup_{c\in\mathcal{R}_t}\mathcal{S}_{tc}.$$

Da die Cluster $s \in S^{tc}$ disjunkt sind, ist es möglich,

$$(\mathcal{C}_{svd}\mathcal{C}_{bk}^2+2\mathcal{C}_{bk})k^2\#\bigcup_{c\in\mathcal{R}_t}\mathcal{S}_{tc} \leq (\mathcal{C}_{svd}\mathcal{C}_{bk}^2+2\mathcal{C}_{bk})k^2\#\mathcal{I}$$

zu erhalten.
Seien nun $t \in \mathcal{T}_{\mathcal{I}} \setminus \mathcal{L}_{\mathcal{I}}$, $c \in \mathcal{R}_t$ und $c' = \mathrm{r}_t(c)$. Das Zusammenkopieren der Matrix $\widehat{A}^{tc}$ benötigt keine arithmetischen Operationen und wird daher nicht mitgezählt. Der Aufwand der anschließenden Singulärwertzerlegung ist geringer als

$$\mathcal{C}_{svd}s_{tc}\min\left\{\sum_{t'\in\mathrm{kind}(t)}k_{t'c'}, s_{tc}\right\}\sum_{t'\in\mathrm{kind}(t)}k_{t'c'} \leq \mathcal{C}_{svd}s_{tc}\mathcal{C}_{kk}^2k^2.$$

Das Kopieren der neuen Transfermatrizen braucht ebenfalls keine arithmetischen Operationen, so dass es nur noch die Berechnung von R_{tc} zu zählen gilt, welche mit

$$2k_{tc}s_{tc}\sum_{t'\in\mathrm{kind}(t)}k_{t'c'} \leq 2s_{tc}\mathcal{C}_{kk}k^2$$

ins Gewicht fällt. Insgesamt beläuft sich der Aufwand in diesem Fall auf

$$\sum_{c\in\mathcal{R}_t}(\mathcal{C}_{svd}\mathcal{C}_{kk}^2+2\mathcal{C}_{kk})k^2 s_{tc} \leq (\mathcal{C}_{svd}\mathcal{C}_{kk}^2+2\mathcal{C}_{kk})k^2\#\mathcal{I}.$$

Für alle Cluster führt dies mit Korollar 3.17 zu

$$\begin{aligned}\sum_{t\in\mathcal{T}_{\mathcal{I}}}&\max\left\{\mathcal{C}_{svd}\mathcal{C}_{kk}^2+2\mathcal{C}_{kk}, \mathcal{C}_{svd}\mathcal{C}_{bk}^2+2\mathcal{C}_{bk}\right\}k^2\#\mathcal{I}\\ &\leq \max\left\{\mathcal{C}_{svd}\mathcal{C}_{kk}^2+2\mathcal{C}_{kk}, \mathcal{C}_{svd}\mathcal{C}_{bk}^2+2\mathcal{C}_{bk}\right\}k^2\#\mathcal{T}_{\mathcal{I}}\#\mathcal{I}\\ &\leq \max\left\{\mathcal{C}_{svd}\mathcal{C}_{kk}^2+2\mathcal{C}_{kk}, \mathcal{C}_{svd}\mathcal{C}_{bk}^2+2\mathcal{C}_{bk}\right\}2\mathcal{C}_{bk}k(\#\mathcal{I})^2.\end{aligned}$$

□[6,*Thm.*17]

Nachdem die richtungsabhängigen Clusterbasen bestimmt sind, gilt es, die Kopplungsmatrizen aufzustellen, was nach (5.2.1) jedoch relativ einfach zu erledigen ist.

Leider liegen die Matrizen der Clusterbasen nur in den Blättern des Clusterbaums direkt vor, so dass für einige Blätter des Blockbaums auch eine eingeschränkte Vorwärtstransformation notwendig wird. Die Matrix $A|_{t\times s}$ wird dazu spaltenweise mit der Clusterbasis multipliziert. Die Vorwärtstransformation ist dabei nicht vollständig durchzuführen, da nicht der ganze Clusterbaum mit allen vorhandenen Richtungen durchlaufen werden muss, sondern nur ein Pfad von Richtungen betrachtet wird.
Der Aufwand für das Entpacken und Multiplizieren einer Clusterbasis für eine Richtung wird im folgenden Lemma abgeschätzt.

Lemma 5.11 (Aufwand Multiplikation der Clusterbasis entlang einer Richtung)
Der Aufwand der Matrix-Multiplikation einer Matrix V_{tc} einer Clusterbasis zum Cluster $t \in \mathcal{T}_\mathcal{I}$ entlang einer Richtung $c \in \mathcal{R}_t$ mit einer passenden Matrix A beträgt

$$mk(\#^\mathfrak{I} t)\mathcal{C}_{ecb},$$

wobei m die Anzahl der Spalten von A und $\mathcal{C}_{ecb} := 4\mathcal{C}_{bk}\max\{\mathcal{C}_{kk}, \mathcal{C}_{bk}\}$ sei.

Beweis: Seien ein Cluster $t \in \mathcal{T}_\mathcal{I}$ sowie ein $c \in \mathcal{R}_t$ gegeben und bezeichne mit $\mathcal{T}_t$ den Teilbaum von $\mathcal{T}_\mathcal{I}$ mit t als Wurzel.
Sei weiter $A \in \mathbb{C}^{k_{tc}\times m}$ eine Matrix mit $m \in \mathbb{N}$ Spalten und interpretiere die Matrix-Multiplikation $V_{tc}A$ als m hintereinander ausgeführte Matrix-Vektor-Multiplikationen für die Spalten von A. Der Aufwand kann ähnlich wie im Beweis des Lemma 3.19 angegeben werden, indem mit $\#\mathcal{R}_{t'}^{eff} = 1$ für alle Cluster t' im Teilbaum $\mathcal{T}_t$ gearbeitet wird. Da nur die Richtung c für die Clusterbasis zum Cluster t betrachtet wird, ist für jedes t' des Clusterbaums $\mathcal{T}_t$ nur eine Richtung nötig. Damit beläuft sich der Aufwand für eine Matrix-Vektor-Multiplikation mit der Clusterbasis zum Clusterbaum $\mathcal{T}_t$ entlang einer Richtung c auf

$$\begin{aligned} 2k^2\max\{\mathcal{C}_{kk}, \mathcal{C}_{bk}\}\sum_{t\in\mathcal{T}_t} 1 &= 2k^2\max\{\mathcal{C}_{kk}, \mathcal{C}_{bk}\}(\#\mathcal{T}_t) \\ &\leq 4\mathcal{C}_{bk}k\max\{\mathcal{C}_{kk}, \mathcal{C}_{bk}\}(\#^\mathfrak{I} t). \end{aligned}$$

Bei m Spalten und mit $\mathcal{C}_{ecb} = 4\mathcal{C}_{bk}\max\{\mathcal{C}_{kk}, \mathcal{C}_{bk}\}$ kann der Gesamtaufwand durch

$$mk(\#^\mathfrak{I} t)\mathcal{C}_{ecb}$$

beschränkt werden. □

Das Lemma 5.11 kann auch für die Multiplikation mit der adjungierten Matrix V_{tc}^* der Clusterbasis verwendet werden, da die Matrix-Matrix-Multiplikation auf die Matrix-Vektor-Multiplikation zurückgeführt wurde und der Aufwand von der Anzahl der Matrixeinträge abhängig ist.

procedure coupling($\mathcal{T}_{\mathcal{I}\times\mathcal{I}}, \mathcal{R}, A, V, W, S$)
  Matrix B
  for $b = (t, s, c) \in \mathcal{L}^+_{\mathcal{I}\times\mathcal{I}}$ **do**
    $B \leftarrow V_{tc}^* A|_{t\times s}$ if necesssary with forward transformation
    $S_b^* \leftarrow W_{sc}^* B^*$ if necesssary with forward transformation
  end for
end procedure

Algorithmus 5.3: Berechnung der Kopplungsmatrizen

Lemma 5.12 (Aufwand Kopplungsmatrizen)
Der Aufwand zur Bestimmung der Kopplungsmatrizen beträgt weniger als

$$2\mathcal{C}_{ecb}k(\#\mathcal{I})^2$$

Operationen.

Beweis: Zum Bestimmen der Kopplungsmatrix sind zwei Matrix-Multiplikationen nötig. Für die Berechnung von $V_{tc}^* A|_{t\times s}$ ist ein Aufwand von $\mathcal{C}_{ecb}(\#^{\mathfrak{I}} t)(\#^{\mathfrak{I}} s)k_{tc}$ nötig. Für die anschließende Multiplikation mit der Spaltenclusterbasis kommt noch ein Aufwand von $\mathcal{C}_{ecb}(\#^{\mathfrak{I}} s)k_{sc}k_{tc}$ hinzu, so dass sich für alle Kopplungsmatrizen zusammen ein Aufwand von

$$\mathcal{C}_{ecb} \sum_{(t,s,c)\in\mathcal{L}^+_{\mathcal{I}\times\mathcal{I}}} (\#^{\mathfrak{I}} t)(\#^{\mathfrak{I}} s)k_{tc} + (\#^{\mathfrak{I}} s)k_{sc}k_{tc}$$

ergibt. Durch den Algorithmus ist sichergestellt, dass $k_{sc} \leq \#^{\mathfrak{I}} s$ gilt, so dass

$$\begin{aligned}
\mathcal{C}_{ecb} \sum_{(t,s,c)\in\mathcal{L}^+_{\mathcal{I}\times\mathcal{I}}} & (\#^{\mathfrak{I}} t)(\#^{\mathfrak{I}} s)k_{tc} + (\#^{\mathfrak{I}} t)k_{sc}k_{tc} \\
&\leq \mathcal{C}_{ecb} \sum_{(t,s,c)\in\mathcal{L}^+_{\mathcal{I}\times\mathcal{I}}} (\#^{\mathfrak{I}} t)(\#^{\mathfrak{I}} s)k_{tc} + (\#^{\mathfrak{I}} t)(\#^{\mathfrak{I}} s)k_{tc} \\
&\leq 2\mathcal{C}_{ecb}k \sum_{(t,s,c)\in\mathcal{L}^+_{\mathcal{I}\times\mathcal{I}}} (\#^{\mathfrak{I}} t)(\#^{\mathfrak{I}} s)
\end{aligned}$$

folgt. Die Blattpartition 2.29 liefert dann

$$2\mathcal{C}_{ecb}k \sum_{(t,s,c)\in\mathcal{L}^+_{\mathcal{I}\times\mathcal{I}}} (\#^{\mathfrak{I}} t)(\#^{\mathfrak{I}} s) \leq 2\mathcal{C}_{ecb}k(\#\mathcal{I})^2.$$

□

Der Aufwand für die Berechnung der Kopplungsmatrizen ist entsprechend in $\mathcal{O}(k(\#\mathcal{I})^2)$.

procedure compression($\mathcal{T}_{\mathcal{I}\times\mathcal{I}}, \mathcal{R}, A, \epsilon, V, W, S$)
  compress_bases($\mathcal{T}_{\mathcal{I}}, \mathcal{R}, \text{wurzel}(\mathcal{T}_{\mathcal{I}}), A, \epsilon, V, R$)
  compress_bases($\mathcal{T}_{\mathcal{I}}, \mathcal{R}, \text{wurzel}(\mathcal{T}_{\mathcal{I}}), A^*, \epsilon, W, R$)
  coupling($\mathcal{T}_{\mathcal{I}\times\mathcal{I}}, \mathcal{R}, A, V, W, S$)
end procedure

Algorithmus 5.4: Kompression einer vollbesetzten Matrix

Theorem 5.13 (Aufwand Kompression)
Der Aufwand der Kompression einer Matrix $A \in \mathbb{C}^{\mathcal{I}\times\mathcal{I}}$ zu einem gegebenen richtungsabhängigen Blockbaum $\mathcal{T}_{\mathcal{I}\times\mathcal{I}}$ ist durch

$$k(\#\mathcal{I})^2 2(\mathcal{C}_{ecb} + \mathcal{C}_{ccb})$$

Operationen beschränkt.

Beweis: Die Abschätzung ergibt sich aus der Addition des Aufwands der richtungsabhängigen Clusterbasis 5.10 (zweifach) mit dem Aufwand für die Kopplungsmatrizen 5.12. □

5.3 Rekompression

Bei der *Rekompression* wird eine vorhandene $\mathcal{RH}^2$-Matrix A durch eine $\mathcal{RH}^2$-Matrix A^{new} approximiert. Ziel dieses zunächst unlogisch erscheinenden Vorhabens ist es, für eine vorher festgelegte Genauigkeit die niedrigsten ausreichenden Ränge zu erreichen und damit die Minimierung der Komplexität zu einer gegebenen Genauigkeit.
Schnelle Ansätze wie die Interpolation liefern konstante Ränge, oftmals sind hohe Ränge aber nur in wenigen Teilmatrizen nötig, so dass viele Teilmatrizen mit deutlich geringerem Aufwand bei gleichbleibender Genauigkeit behandelt werden könnten. Ein algebraischer Ansatz unter Einsatz der Singulärwertzerlegung führt zu den bestmöglichen Kompressionsraten in Bezug auf die Spektral- und Frobeniusnorm, ist jedoch zeitintensiv. Eine Kombination aus einem schnellen Ansatz zum Aufstellen einer $\mathcal{RH}^2$-Matrix und einer Kompression durch Singulärwertzerlegungen erlaubt es, fast optimale Kompressionsraten in deutlich geringerer Zeit zu erreichen. Der Algorithmus sowie eine Aufwandsabschätzung der Rekompression wurden in Zusammenarbeit mit Herrn Börm in [10] veröffentlicht und werden im Folgenden umfangreicher ausgeführt.

Hierfür wird es erneut notwendig, zu einem Cluster $t \in \mathcal{T}_{\mathcal{I}}$ und einer festen Richtung $c \in \mathcal{R}_t$, die Cluster $s \in \text{row}^+(t)$ zu betrachten, die mit t einen zulässigen Block bilden.

Definiere dazu eine für Richtungsabhängigkeiten eingeschränkte Version der Menge row^+

$$\mathrm{row}_c^+(t) := \left\{ s \in \mathrm{row}^+(t) \mid (t,s,c) \in \mathcal{L}^+_{\mathcal{I}\times\mathcal{I}} \right\}$$

und entsprechend

$$\mathrm{col}_c^+(s) := \left\{ t \in \mathrm{col}^+(s) \mid (t,s,c) \in \mathcal{L}^+_{\mathcal{I}\times\mathcal{I}} \right\}.$$

Im Gegensatz zur direkten Kompression liegen aufgrund des hierarchischen Aufbaus nicht mehr alle Teilmatrizen direkt vor, um die Matrix A^{tc} aufstellen zu können. Aus diesem Grund wird zunächst ein passendes Äquivalent zur Matrix A^{tc} in Form von *totalen Clusterbasen* aufgestellt.
Eine totale Clusterbasis enthält nicht nur die Informationen zu ihrem Cluster t, sondern auch zu zulässigen Blöcken, die zu Vorfahren gehören, also zu Blöcken, die über die Schachtelungseigenschaft der Clusterbasen aus den Informationen zum Cluster t rekonstruiert werden müssen. Erst dies ermöglicht das Reduzieren auf das Relevante und damit einhergehend die Reduktion des Rangs. Dazu werden auch Transfermatrizen benötigt, die über mehrere Stufen hinweg arbeiten können.

Definition 5.14 (Weitreichende Transfermatrix)
Für alle Cluster $t \in \mathcal{T}_\mathcal{I}, t^+ \in \mathrm{vor}(t)$ und Richtungen $c \in \mathcal{R}_t, c^+ \in \mathrm{vor}_t^{t^+}(c)$ ist die weitreichende Transfermatrix E_{tc,t^+c^+} definiert durch

$$E_{tc,t^+c^+} := \begin{cases} E_{tc,\tilde{t}\tilde{c}} E_{\tilde{t}c^+} & \text{falls } \tilde{t} \in \mathrm{kind}(t^+) \text{ existiert mit } t \in \mathrm{nac}(\tilde{t}), \tilde{c} = \mathrm{r}_{t^+}(c^+), \\ Id & \text{sonst}. \end{cases}$$

Definition 5.15 (Totale richtungsabhängige Clusterbasis)
Sei eine $\mathcal{RH}^2$-Matrix $A \in \mathbb{C}^{\mathcal{I}\times\mathcal{I}}$ mit richtungsabhängigen Clusterbasen $\{V_{tc}\}_{\substack{t\in\mathcal{T}_\mathcal{I}\\ c\in\mathcal{R}_t}}$, $\{W_{sc}\}_{\substack{s\in\mathcal{T}_\mathcal{I}\\ c\in\mathcal{R}_s}}$ und einer Familie von Richtungsmengen $\mathcal{R}$ gegeben. Bezeichne die Familie $\{A_{tc}\}_{\substack{t\in\mathcal{T}_\mathcal{I}\\ c\in\mathcal{R}_t}}$, die durch

$$A_{tc} := V_{tc} X_{tc}^* \in \mathbb{C}^{\mathcal{I}\times\mathcal{I}}_{\mathfrak{I}_{t\times\mathcal{S}_{tc}}} \qquad \text{für alle } t \in \mathcal{T}_\mathcal{I}, c \in \mathcal{R}_t \text{ mit } k_{tc} \neq 0 \tag{5.3.1}$$

mit dem Gewicht

$$\mathbb{C}^{\mathcal{I}\times k_{tc}}_{\mathcal{S}_{tc}\times k_{tc}} \ni X_{tc} := \sum_{\substack{t^+\in\mathrm{vor}(t)\\ c^+\in\mathrm{vor}_t^{t^+}(c)}} \sum_{s\in\mathrm{row}^+_{c^+}(t^+)} W_{sc^+} S^*_{(t^+,s,c^+)} E^*_{tc,t^+c^+} \tag{5.3.2}$$

gegeben ist, als totale richtungsabhängige Clusterbasis *zum Cluster t und zur Richtung c.*

Bemerkung 30 : *Die verwendeten Spaltencluster in der Definition des Gewichts* (5.3.2) *entsprechen denen in der Menge S^{tc}* (5.2.4)*, jedoch ist es für die Betrachtung des Algorithmus einfacher, die ausführliche Formulierung zu verwenden.*

Bemerkung 31 : *Eine totale richtungsabhängige Clusterbasis für einen Spaltencluster $s \in \mathcal{T}_\mathcal{I}$ lässt sich auf die gleiche Weise durch Betrachten der adjungierten Matrix A^* erhalten, es folgt*

$$A_{sc} = W_{sc} \sum_{\substack{s^+ \in \mathrm{vor}(s) \\ c^+ \in \mathrm{vor}_s^{s^+}(c)}} \sum_{t \in \mathrm{col}_{c^+}^+(s^+)} E_{sc,s^+c^+} S^*_{(t,s^+,c^+)} V^*_{tc^+}.$$

Bei der totalen richtungsabhängigen Clusterbasis zu einem Cluster $t \in \mathcal{T}_\mathcal{I}$ und einer Richtung $c \in \mathcal{R}_t$ handelt es sich um das Produkt der richtungsabhängigen Matrix V_{tc} mit einem Gewicht, dies ermöglicht es, alle zulässigen Blöcke, die mit t oder einem Vorfahren t^+ von t und der entsprechenden Vorfahrenrichtung c^+ gebildet werden, auf einmal zu betrachten.

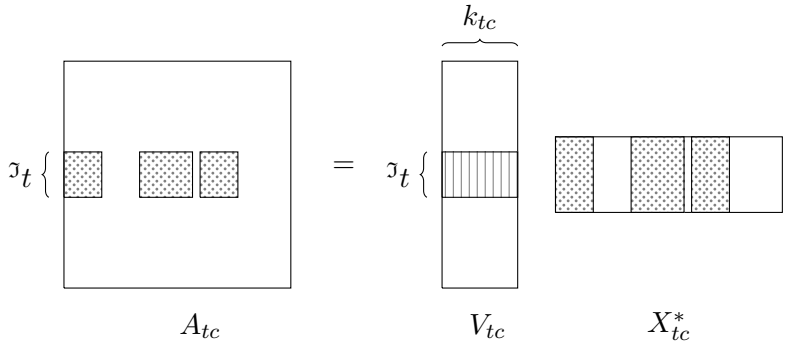


Abbildung 5.3: Darstellung der totalen richtungsabhängigen Clusterbasis

In der Abbildung 5.3 ist gezeigt, wie sich die totale Clusterbasis A_{tc} als Produkt aus der Matrix V_{tc} und der Gewichtsmatrix X_{tc} zusammensetzt. Dabei sind alle Cluster $s \in S^{tc}$ in A_{tc} und X_{tc} hervorgehoben.
Die Reduktion des Rangs soll, wie bereits erwähnt, mit Hilfe einer Singulärwertzerlegung (siehe Definition und Lemma 5.7) geschehen. Dabei bleibt zu bedenken, dass die Singulärwertzerlegung sehr aufwendig ist (5.2.2). Aus diesem Grund empfiehlt es sich, für die praktische Berechnung die Matrix A_{tc} zu kondensieren und Fallunterscheidungen einzuführen, da kleinere Matrizen einen niedrigeren Aufwand für die Singulärwertzerlegung bedeuten. Zusätzlich kann für die Darstellung des Gewichts für ein Cluster $t \in \mathcal{T}_\mathcal{I}$ mit Richtung $c \in \mathcal{R}_t$ eine rekursive Formulierung gefunden werden, welche den Rechenaufwand nochmals reduziert.
Dabei reicht es vollkommen aus, die totale richtungsabhängige Clusterbasis nur für die Zeilencluster zu betrachten, da der Fall der Spaltenclusterbasis, wie in Bemerkung 31 angedeutet, erneut über Adjungieren der Matrix A behandelt werden kann.
Da auch der praktische Algorithmus zunächst die Gewichte X_{tc} bestimmt, soll hier ebenfalls mit der rekursiven Formulierung für die Gewichte begonnen werden.

Sei $t \in \mathcal{T}_\mathcal{I}$ ein Cluster, welcher an mindestens einem zulässigen Block mit Richtung $c \in \mathcal{R}_t$ beteiligt ist. Existiert kein zulässiger Block (t^+, s^+, c^+), der mit einem Vorfahren $t^+ \in \mathrm{vor}(t)$ und $c^+ \in \mathrm{vor}_t^{t^+}(c)$ gebildet wird, so vereinfacht sich (5.3.2) zu

$$X_{tc} = \sum_{s \in \mathrm{row}_c^+(t)} W_{sc} S^*_{(t,s,c)}.$$

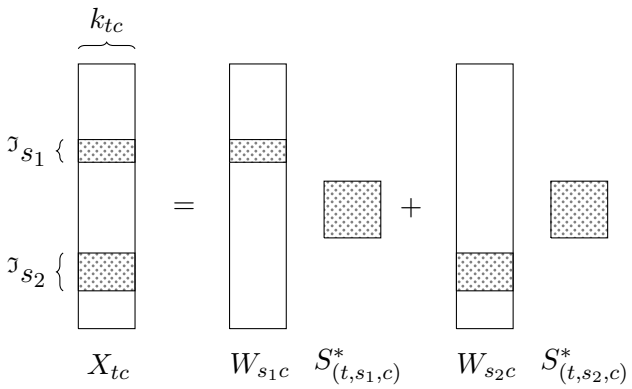


Abbildung 5.4: Beispiel für die Gewichtsmatrix

Ein Beispiel für die Gewichtsmatrix bei zwei zulässigen Blöcken, falls keine zulässigen Vorfahren existieren, ist in Abbildung 5.4 dargestellt.
Angenommen für t und alle Richtungen $c \in \mathcal{R}_t$ ist X_{tc} schon bestimmt und t hat ein Kind $t' \in \mathrm{kind}(t)$ mit $c' = \mathrm{r}_t(c) \in \mathcal{R}_{t'}$. Dann kann für alle $t^+ \in \mathrm{vor}(t)$ mit $c^+ \in \mathrm{vor}_t^{t^+}(c)$ die weitreichende Transfermatrix $E_{t'c',t^+c^+}$ aus Definition 5.14 auch mit

$$E_{t'c',t^+c^+} = E_{t'c} E_{tc,t^+c^+}$$

geschrieben werden. Somit kann das Gewicht des Kindes t' von t durch Herausziehen der Summanden, die auf der Stufe $\mathrm{stufe}(t')$ sind, umgeformt werden

$$\begin{aligned} X_{t'c'} &= \sum_{\substack{t^+ \in \mathrm{vor}(t') \\ c^+ \in \mathrm{vor}_{t'}^{t^+}(c')}} \sum_{s \in \mathrm{row}_{c^+}^+(t^+)} W_{sc^+} S^*_{(t^+,s,c^+)} E^*_{t'c',t^+c^+} \\ &= \sum_{s \in \mathrm{row}_{c'}^+(t')} W_{sc'} S^*_{(t',s,c')} + \sum_{c \in \mathrm{vor}_{t'}^t(c')} \sum_{\substack{t^+ \in \mathrm{vor}(t) \\ c^+ \in \mathrm{vor}_t^{t^+}(c)}} \sum_{s \in \mathrm{row}_{c^+}^+(t^+)} W_{sc^+} S^*_{(t^+,s,c^+)} E^*_{tc,t^+c^+} E^*_{t'c}. \end{aligned}$$

In der zweiten Summe steckt X_{tc}, so dass die Gleichung deutlich kürzer mit

$$X_{t'c'} = \sum_{s \in \mathrm{row}_{c'}^+(t')} W_{sc'} S^*_{(t',s,c')} + \sum_{c \in \mathrm{vor}_{t'}^t(c')} X_{tc} E^*_{t'c}$$

geschrieben werden kann. Auf diese Weise ergibt sich die rekursive Formulierung mit

$$X_{tc} = \begin{cases} \displaystyle\sum_{c^+ \in \operatorname{vor}_t^{t^+}(c)} X_{t^+c^+} E_{tc^+}^* + \sum_{s \in \operatorname{row}_c^+(t)} W_{sc} S_{(t,s,c)}^* & \text{falls ein } t^+ \in \mathcal{T}_\mathcal{I} \text{ mit} \\ & t \in \operatorname{kind}(t^+) \text{ existiert,} \\ \displaystyle\sum_{s \in \operatorname{row}_c^+(t)} W_{sc} S_{(t,s,c)}^* & \text{sonst,} \end{cases}$$

wobei die Matrizen $X_{t^+c^+}$ leer sind, falls weder t^+ noch ein Vorfahre von t^+ an einem zulässigen Block beteiligt sind.
Durch Orthogonalisieren der Clusterbasis $\{W_{sc}\}_{\substack{s \in \mathcal{T}_\mathcal{I} \\ c \in \mathcal{R}_s}}$ kann das Gewicht weiter umgeschrieben und der Aufwand reduziert werden. Betrachte zu Beginn den Fall ohne Vorfahren genauer. Dann existieren für alle $s \in \operatorname{row}_c^+(t)$ mit $W_{sc} \in \mathbb{C}^{\mathcal{I} \times k_{sc}}_{\mathfrak{I} s \times k_{sc}}$ ein $q_{sc} = \min\left\{\#^\mathfrak{I} s, k_{sc}\right\}$, eine isometrische Matrix $Q_{sc} \in \mathbb{C}^{\mathcal{I} \times q_{sc}}_{\mathfrak{I} s \times q_{sc}}$ sowie eine obere Dreiecksmatrix $R_{sc} \in \mathbb{C}^{q_{sc} \times k_{sc}}$ mit $W_{sc} = Q_{sc} R_{sc}$. Mit $\operatorname{row}_c^+(t) = \{s_1, \ldots, s_\sigma\}$ und $m = \sum_{i=1}^{\sigma} q_{s_i c}$ ergibt sich die Darstellung

$$X_{tc} = \sum_{s \in \operatorname{row}_c^+(t)} Q_{sc} R_{sc} S_{(t,s,c)}^* = \underbrace{(Q_{s_1 c} \ldots Q_{s_\sigma c})}_{=: \hat{U}_{tc} \in \mathbb{C}^{\mathcal{I} \times m}_{\mathcal{S}_{tc} \times m}} \underbrace{\begin{pmatrix} R_{s_1 c} S_{(t,s_1,c)}^* \\ \vdots \\ R_{s_\sigma c} S_{(t,s_\sigma,c)}^* \end{pmatrix}}_{=: Y_{tc} \in \mathbb{C}^{m \times k_{tc}}}.$$

Auch bei $\hat{U}_{tc}$ handelt es sich um eine isometrische Matrix, da sie als Blockmatrix aus isometrischen Matrizen zusammengesetzt wird und die vorkommenden Spaltenclusterpartner aus $\operatorname{row}_c^+(t)$ auf Grund der Blattpartition des Blockbaums leere Schnitte haben. Die Matrix Y_{tc} kann ebenfalls orthogonalisiert werden. Für $\widetilde{m} = \min\{m, k_{tc}\}$ existieren eine isometrische Matrix $P_{tc} \in \mathbb{C}^{m \times \widetilde{m}}$ und eine obere Dreiecksmatrix $Z_{tc} \in \mathbb{C}^{\widetilde{m} \times k_{tc}}$ mit

$$Y_{tc} = P_{tc} Z_{tc}.$$

Als Produkt isometrischer Matrizen ist $\hat{U}_{tc} P_{tc} =: U_{tc} \in \mathbb{C}^{\mathcal{I} \times \widetilde{m}}_{\mathcal{S}_{tc} \times \widetilde{m}}$ ebenfalls isometrisch. Insgesamt wird eine Darstellung für X_{tc} mit einer isometrischen Matrix U_{tc} und einer oberen Dreiecksmatrix Z_{tc} durch

$$X_{tc} = \hat{U}_{tc} Y_{tc} = \hat{U}_{tc} P_{tc} Z_{tc} = U_{tc} Z_{tc}$$

erhalten.
Solch eine Darstellung kann auch im komplizierten Fall generiert werden, nehme dazu an, dass für den Elterncluster t^+ von t schon eine entsprechende Darstellung bestimmt wurde. Erneut orthogonalisiere die Spaltenclusterbasis, es existieren also für alle $s \in \operatorname{row}_c^+(t)$ ein $q_{sc} = \min\left\{\#^\mathfrak{I} s, k_{sc}\right\}$ sowie eine isometrische Matrix $Q_{sc} \in \mathbb{C}^{\mathcal{I} \times q_{sc}}_{\mathfrak{I} s \times q_{sc}}$ und eine obere

Dreiecksmatrix $R_{sc} \in \mathbb{C}^{q_{sc} \times k_{sc}}$ mit $W_{sc} = Q_{sc} R_{sc}$. Nutze weiterhin, dass für das Gewicht $X_{t^+c^+} \in \mathbb{C}^{\mathcal{I} \times k_{t^+c^+}}_{\mathcal{S}_{t^+c^+} \times k_{t^+c^+}}$ dann zu einem $\widetilde{m}_{(t^+)}$ eine isometrische Matrix $U_{t^+c^+} \in \mathbb{C}^{\mathcal{I} \times \widetilde{m}_{(t^+)}}_{\mathcal{S}_{t^+c^+} \times \widetilde{m}_{(t^+)}}$ und eine Matrix $Z_{t^+c^+} \in \mathbb{C}^{\widetilde{m}_{(t^+)} \times k_{t^+c^+}}$ mit $X_{t^+c^+} = U_{t^+c^+} Z_{t^+c^+}$ existieren. Für $m := \sum_{c^+ \in \operatorname{vor}_t^{t^+}(c)} \widetilde{m}_{(t^+)}(c^+) + \sum_{s \in \operatorname{row}_c^+(t)} q_{sc}$ und $\operatorname{row}_c^+(t) = \{s_1, \ldots, s_\sigma\}$ sowie Vorfahrenrichtungen $\operatorname{vor}_t^{t^+}(c) = \{c_1^+, \ldots, c_r^+\}$ folgt

$$\begin{aligned}
X_{tc} &= \sum_{c^+ \in \operatorname{vor}_t^{t^+}(c)} X_{t^+c^+} E^*_{tc^+} + \sum_{s \in \operatorname{row}_c^+(t)} W_{sc} S^*_{(t,s,c)} \\
&= \sum_{c^+ \in \operatorname{vor}_t^{t^+}(c)} U_{t^+c^+} Z_{t^+c^+} E^*_{tc^+} + \sum_{s \in \operatorname{row}_c^+(t)} Q_{sc} R_{sc} S^*_{(t,s,c)} \\
&= \underbrace{\begin{pmatrix} U_{t^+c_1^+} & \cdots & U_{t^+c_r^+} & Q_{s_1 c} & \cdots & Q_{s_\sigma c} \end{pmatrix}}_{=:\hat{U}_{tc} \in \mathbb{C}^{\mathcal{I} \times m}_{\mathcal{S}_{tc} \times m}} \underbrace{\begin{pmatrix} Z_{t^+c_1^+} E^*_{tc_1^+} \\ \vdots \\ Z_{t^+c_r^+} E^*_{tc_r^+} \\ R_{s_1 c} S^*_{(t,s_1,c)} \\ \vdots \\ R_{s_\sigma c} S^*_{(t,s_\sigma,c)} \end{pmatrix}}_{=:Y_{tc} \in \mathbb{C}^{m \times k_{tc}}}.
\end{aligned}$$

Auch in diesem Fall ist $\hat{U}_{tc}$ als Blockmatrix isometrischer Matrizen eine isometrische Matrix, da erneut auf Grund der Blattpartition des Blockbaums die Schnitte der auftretenden Spaltenclusterpartner leer sind.
Nach der Orthogonalisierung von Y_{tc} mit $\widetilde{m} = \min\{m, k_{tc}\}$ existieren eine isometrische Matrix $P_{tc} \in \mathbb{C}^{m \times \widetilde{m}}$ und eine obere Dreiecksmatrix $Z_{tc} \in \mathbb{C}^{\widetilde{m} \times k_{tc}}$ mit

$$Y_{tc} = P_{tc} Z_{tc}.$$

Setze $U_{tc} := \hat{U}_{tc} P_{tc} \in \mathbb{C}^{\mathcal{I} \times \widetilde{m}}_{\mathcal{S}_{tc} \times \widetilde{m}}$ und gewinne so die gesuchte Darstellung

$$X_{tc} = U_{tc} Z_{tc}.$$

Um diese Faktorisierung zu erhalten, wurde viel Aufwand betrieben. Dies lohnt sich jedoch auch, da Z_{tc} meist weniger Speicher benötigt als Y_{tc} und somit eine effizientere Darstellung der Gewichte zugänglich ist. Außerdem wird von der Matrix X_{tc} nur das Z_{tc} für die Singulärwertzerlegung verwendet, da U_{tc} nur auf die rechten Singulärvektoren Einfluss nimmt, so dass sich auch hier eine Einsparung in der Rechenzeit ergibt.

Falls anstatt eines Zeilenclusters t ein Spaltencluster s betrachtet wird, ist der Ausgangspunkt die adjungierte Matrix A^*. Mit analoger Vorgehensweise ergibt sich für

$$A_{sc} = W_{sc} X_{sc}^* = W_{sc} \sum_{\substack{s^+ \in \mathrm{vor}(s) \\ c^+ \in \mathrm{vor}_s^{s^+}(c)}} \sum_{t \in \mathrm{col}_{c^+}^+(s^+)} E_{sc,s^+c^+} S_{(t,s^+,c^+)}^* V_{tc^+}^*$$

die rekursive Formulierung

$$X_{sc} = \begin{cases} \displaystyle\sum_{c^+ \in \mathrm{vor}_s^{s^+}(c)} X_{s^+c^+} E_{sc^+}^* + \sum_{t \in \mathrm{col}_c^+(s)} V_{tc} S_{(t,s,c)} & \text{falls ein } s^+ \in \mathcal{T}_\mathcal{I} \text{ mit} \\ & s \in \mathrm{kind}(s^+) \text{ existiert,} \\ \displaystyle\sum_{t \in \mathrm{col}_c^+(s)} V_{tc} S_{(t,s,c)} & \text{sonst.} \end{cases}$$

Für die vorgenommenen Umformungen wurde davon ausgegangen, dass die betrachteten Clusterbasen orthogonalisiert werden müssen. An sich braucht die Clusterbasis selbst nicht isometrisch zu sein, da sie im Verlauf der Rekompression ohnehin neu berechnet wird. Entsprechend reicht es vollkommen aus, die Matrizen für den Basiswechsel zu bestimmen und zu speichern. Es gilt dann nur, zu bedenken, dass es neben dem ursprünglichen Rang k_{tc} noch den neuen Rang k_{tc}^{new}, welcher aus der Orthogonalisierung stammt, gibt.
Falls von vornherein mit einer isometrischen Clusterbasis gearbeitet wird, lässt sich der Algorithmus leicht dahingehend modifizieren. Die Multiplikation mit der jeweiligen anderen Clusterbasis entfällt, stattdessen wird nur die adjungierte Kopplungsmatrix kopiert. Zusätzlich braucht nicht zwischen k_{tc}^{new} und k_{tc} unterschieden werden.

Aus Gründen der Übersichtlichkeit wird der Algorithmus für die Gewichte hier nur für den Fall der Zeilenclusterbasis angegeben. Dabei sind die den Basiswechsel zur isometrischen Clusterbasis beschreibenden Matrizen in der Familie von Matrizen R enthalten. Werden die Rollen der Clusterbasen und das Adjungieren der $\mathcal{RH}^2$- Matrix bedacht, läuft er für die Spaltencluster analog ab. Bezeichne mit S die Menge aller Kopplungsmatrizen der betrachteten Matrix A und speichere die berechneten Gewichte in Z, einer Familie von Matrizen.

Lemma 5.16 (Komplexität der Gewichte)
Die Anzahl der Operationen beim Bestimmen der Gewichtsmatrizen einer richtungsabhängigen Clusterbasis $\{V_{tc}\}_{\substack{t \in \mathcal{T}_\mathcal{I} \\ c \in \mathcal{R}_t}}$ einer $\mathcal{RH}^2$-Matrix ist beschränkt durch

$$k^3 \mathcal{C}_{wei} \left((\mathcal{C}_{sk} + \mathcal{C}_{kk}) \left(\#\mathcal{T}_\mathcal{I} + p_\mathcal{I} \kappa^2 \mathcal{C}_w \right) + 1 \right), \tag{5.3.3}$$

wobei die Konstanten $\mathcal{C}_{wei}$ und $\mathcal{C}_w$ durch

$$\mathcal{C}_{wei} := (2 + \mathcal{C}_{qr}) \qquad \textit{und} \qquad \mathcal{C}_w := \max \left\{ \mathcal{C}_{kk} \mathcal{C}_{Ck} \mathcal{C}_{uk} \mathcal{C}_{mk}^2 |\Gamma|, \mathcal{C}_{tk} \right\}$$

gegeben sind.

procedure row_weights($\mathcal{T}_{\mathcal{I}\times\mathcal{I}}, \mathcal{R}, t, V, Z, R, S$)
Matrices T, Q, P, int m, n
for all $c \in \mathcal{R}_t$ **do**
$Z_{tc} \leftarrow 0$
for all $s \in \mathrm{row}_c^+(t)$ **do**
$n \leftarrow m + k_{sc}^{new}$, $\mathbb{C}^{n\times k_{tc}} \ni T \leftarrow \begin{pmatrix} Z_{tc} \\ R_{sc}S_b^* \end{pmatrix}$, $\triangleright\ m\widehat{=}$ rows Z_{tc}
$m \leftarrow min\{n, k_{tc}\}$, compute QR decomposition $T = QP$
$\mathbb{C}^{m\times k_{tc}} \ni Z_{tc} \leftarrow P$
end for
if there is a parent cluster t^+ of t **then**
$\mathrm{vor}_t^{t^+}(c) = \{c_1^+, \ldots, c_\tau^+\}$, $n \leftarrow m + \sum_{i=1}^{\tau} m_{(t^+)}(c_i^+)$ $\triangleright\ m\widehat{=}$ rows Z_{tc}
$\mathbb{C}^{n\times k_{tc}} \ni T \leftarrow \begin{pmatrix} Z_{t^+c_1^+}E^*_{tc_1^+} \\ \vdots \\ Z_{t^+c_\tau^+}E^*_{tc_\tau^+} \\ Z_{tc} \end{pmatrix}$, $m \leftarrow min\{n, k_{tc}\}$
Compute QR decomposition $T = QP$, $\mathbb{C}^{m\times k_{tc}} \ni Z_{tc} \leftarrow P$
end if
end for
for all $t' \in \mathrm{kind}(t)$ **do**
row_weights($\mathcal{T}_{\mathcal{I}\times\mathcal{I}}, \mathcal{R}, t', V, Z, R, S$)
end for
end procedure

Algorithmus 5.5: Bestimmen der Gewichte

Beweis: Seien ein $t \in \mathcal{T}_\mathcal{I}$ und eine Richtung $c \in \mathcal{R}_t^{eff}$ gegeben.
1. Fall Es gibt keinen Vorfahren von t, der in einem zulässigen Blatt auftritt, t selbst ist aber an mindestens einem mit Richtung c beteiligt. Dann gilt es, die Multiplikation der oberen Dreiecksmatrizen R mit den Kopplungsmatrizen S abzuschätzen, was zunächst auf

$$\sum_{s \in \operatorname{row}_c^+(t)} 2q_{sc}k_{tc}k_{sc} \leq \sum_{s \in \operatorname{row}_c^+(t)} 2k_{tc}k_{sc}^2 \leq 2k^3 \sum_{s \in \operatorname{row}_c^+(t)} 1 = 2k^3 \# \operatorname{row}_c^+(t)$$

führt.
2. Fall Der Elterncluster t^+ ist an zulässigen Blättern beteiligt oder t^+ besitzt Vorfahren, die zulässige Blätter bilden. Dann gilt es, die Multiplikation der Transfermatrizen mit den Matrizen $Z_{t^+c^+}$ abzuschätzen, dafür ergibt sich ein Aufwand von

$$\sum_{\substack{c^+ \in \operatorname{vor}_t^{t^+}(c) \\ c^+ \in \mathcal{R}_{t^+}^{eff}}} 2\widetilde{m}_{(t^+)}(c^+)k_{t^+c^+}k_{tc}.$$

Die Einschränkung der Summe auf Richtungen $c^+ \in \operatorname{vor}_t^{t^+}(c)$, die ebenfalls in $\mathcal{R}_{t^+}^{eff}$ sind, ist möglich, da eine Vorfahrenrichtung nur dann in Betracht gezogen werden muss, wenn sie überhaupt aktiv mit dem Cluster t^+ zusammen verwendet wird. Da für alle Elternrichtungen c^+ gilt, dass $\widetilde{m}_{(t^+)}$ kleiner gleich dem Minimum von $m_{(t^+)}$ und $k_{t^+c^+}$ ist, kann es auch gegen k abgeschätzt werden.
Damit beläuft sich der Aufwand maximal auf

$$\begin{aligned}\sum_{s \in \operatorname{row}_c^+(t)} 2q_{sc}k_{tc}k_{sc} + \sum_{\substack{c^+ \in \operatorname{vor}_t^{t^+}(c) \\ c^+ \in \mathcal{R}_{t^+}^{eff}}} 2\widetilde{m}_{(t^+)}(c^+)k_{t^+c^+}k_{tc} \\ \leq 2k^3 \left(\# \operatorname{row}_c^+(t) + \sum_{\substack{c^+ \in \operatorname{vor}_t^{t^+}(c) \\ c^+ \in \mathcal{R}_{t^+}^{eff}}} 1 \right),\end{aligned}$$

wenn sowohl Vorfahren als auch der Cluster selbst an zulässigen Blöcken beteiligt sind.
Die anschließende QR-Zerlegung der Matrix Y_{tc} hat einen Aufwand von

$$C_{qr} m k_{tc} \min\{m, k_{tc}\}.$$

Dieser kann mit Hilfe von

$$m = \sum_{s \in \operatorname{row}_c^+(t)} q_{sc} + \sum_{\substack{c^+ \in \operatorname{vor}_t^{t^+}(c) \\ c^+ \in \mathcal{R}_{t^+}^{eff}}} \widetilde{m}_{(t^+)}(c^+) \leq k \left(\# \operatorname{row}_c^+(t) + \sum_{\substack{c^+ \in \operatorname{vor}_t^{t^+}(c) \\ c^+ \in \mathcal{R}_{t^+}^{eff}}} 1 \right)$$

durch

$$k^3\mathcal{C}_{qr}\left(\#\operatorname{row}_c^+(t) + \sum_{\substack{c^+\in\operatorname{vor}_t^{t^+}(c)\\ c^+\in\mathcal{R}_{t^+}^{eff}}} 1\right)$$

beschränkt werden.
Für den Gesamtaufwand des richtungsabhängigen Clusterbaums ergibt sich damit ein Aufwand von

$$\sum_{t\in\mathcal{T}_\mathcal{I}}\sum_{c\in\mathcal{R}_t^{eff}} k^3\left(\#\operatorname{row}_c^+(t) + \sum_{\substack{c^+\in\operatorname{vor}_t^{t^+}(c)\\ c^+\in\mathcal{R}_{t^+}^{eff}}} 1\right)(\mathcal{C}_{qr}+2)$$

$$\leq k^3(2+\mathcal{C}_{qr})\sum_{t\in\mathcal{T}_\mathcal{I}}\sum_{c\in\mathcal{R}_t^{eff}}\left(\#\operatorname{row}_c^+(t) + \sum_{\substack{c^+\in\operatorname{vor}_t^{t^+}(c)\\ c^+\in\mathcal{R}_{t^+}^{eff}}} 1\right).$$

Die erste Doppelsumme kann mit Lemma 3.15 abgeschätzt werden. Bei der verbleibenden Dreifachsumme, können zunächst die inneren beiden Summen mit

$$\sum_{c\in\mathcal{R}_t^{eff}}\sum_{\substack{c^+\in\operatorname{vor}_t^{t^+}(c)\\ c^+\in\mathcal{R}_{t^+}^{eff}}} 1 \leq \#\mathcal{R}_{t^+}^{eff}$$

beschränken werden. Wenn dann zusätzlich die Summe über alle Cluster betrachtet wird, gilt

$$\sum_{t\in\mathcal{T}_\mathcal{I}}\#\mathcal{R}_{t^+}^{eff} = \sum_{\ell=0}^{p_\mathcal{I}}\sum_{t\in\mathcal{T}_\mathcal{I}^\ell}\#\mathcal{R}_{t^+}^{eff} \leq \mathcal{C}_{kk}\sum_{\ell=0}^{p_\mathcal{I}-1}\sum_{t^+\in\mathcal{T}_\mathcal{I}^\ell}\#\mathcal{R}_{t^+}^{eff}.$$

Da für jedes Kindercluster die effektiven Richtungen des Elternclusters betrachtet werden, und dies kann auf dieselbe Weise wie in Lemma 3.14 mit

$$\mathcal{C}_{kk}\sum_{\ell=0}^{p_\mathcal{I}-1}\sum_{t^+\in\mathcal{T}_\mathcal{I}^\ell}\#\mathcal{R}_{t^+}^{eff} \leq \mathcal{C}_{kk}\left(\#\mathcal{T}_\mathcal{I} + p_\mathcal{I}\kappa^2\mathcal{C}_{tk}\right)$$

abgeschätzt werden. Setze $\mathcal{C}_w := \max\left\{\mathcal{C}_{kk}\mathcal{C}_{Ck}\mathcal{C}_{uk}\mathcal{C}_{mk}^2|\Gamma|, \mathcal{C}_{tk}\right\}$ und erhalte so

$$
\begin{aligned}
k^3(2+\mathcal{C}_{qr}) \sum_{t\in\mathcal{T}_\mathcal{I}} \sum_{c\in\mathcal{R}_t^{eff}} \left(\#\operatorname{row}_c^+(t) + \sum_{\substack{c^+\in \operatorname{vor}_t^{t^+}(c) \\ c^+\in\mathcal{R}_{t^+}^{eff}}} 1 \right) \\
k^3(2+\mathcal{C}_{qr}) \left(\mathcal{C}_{sk}\left(\#\mathcal{T}_\mathcal{I} + p_\mathcal{I}\kappa^2\mathcal{C}_{kk}\mathcal{C}_{Ck}\mathcal{C}_{uk}\mathcal{C}_{mk}^2|\Gamma|\right) + 1 + \sum_{t\in\mathcal{T}_\mathcal{I}} \#\mathcal{R}_{t^+}^{eff} \right) \\
\leq k^3(2+\mathcal{C}_{qr}) \left(\mathcal{C}_{sk}\left(\#\mathcal{T}_\mathcal{I} + p_\mathcal{I}\kappa^2\mathcal{C}_w\right) + 1 + \mathcal{C}_{kk}\left(\#\mathcal{T}_\mathcal{I} + p_\mathcal{I}\kappa^2\mathcal{C}_{tk}\right)\right) \\
\leq k^3(2+\mathcal{C}_{qr}) \left(\left(\mathcal{C}_{sk}+\mathcal{C}_{kk}\right)\left(\#\mathcal{T}_\mathcal{I} + p_\mathcal{I}\kappa^2\mathcal{C}_w\right) + 1\right).
\end{aligned}
$$

□

Der Aufwand für die Gewichte liegt mit Korollar 3.17 in $\mathcal{O}(k^2(\#\mathcal{I} + k\kappa^2\log_2(\#\mathcal{I})))$.

Bemerkung 32 (Ersparnis isometrische Clusterbasen): *Werden die Gewichte für eine $\mathcal{RH}^2$-Matrix mit isometrischen Clusterbasen bestimmt, entfällt entsprechend der Aufwand von $2k^3\#\operatorname{row}_c^+(t)$ für die Multiplikation der Kopplungsmatrizen mit den Dreiecksmatrizen in R. Im Gesamtaufwand macht sich dies durch eine Konstante $\mathcal{C}_{qr}$ statt $\mathcal{C}_{wei} = 2 + \mathcal{C}_{qr}$ bemerkbar.*

Nachdem die Bestimmung der Gewichte abgeschlossen ist, kann mit ihrer Hilfe die totale richtungsabhängige Clusterbasis noch einmal komprimiert werden. Da ausschließlich die linken Singulärvektoren von Interesse sind und die rechten vernachlässigt werden können, kann die Multiplikation mit der isometrischen Matrix U_{tc} von rechts weggelassen werden, demnach nutze eine kondensierte Form

$$A_{tc}^c = V_{tc}Z_{tc}^*.$$

Zur Reduktion des Rangs verwende eine Singulärwertzerlegung der kondensierten Matrix $A_{tc}^c \in \mathbb{C}^{\mathcal{I}\times\widetilde{m}}_{\mathfrak{I}t\times\widetilde{m}}$. Für $q \leq \min\left\{\#\mathfrak{I}t, \widetilde{m}\right\}$ existieren zwei isometrische Matrizen $U \in \mathbb{C}^{\mathcal{I}\times q}_{\mathfrak{I}t\times q}$ und $O \in \mathbb{C}^{\widetilde{m}\times q}$ mit

$$A_{tc}^c = U diag(\sigma_1, \ldots, \sigma_q) O^*,$$

wobei $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_q$ gilt. Mit Hilfe der Fehlerschranken für die abgeschnittene Singulärwertzerlegung (5.2.3) kann der minimale benötigte Rang bestimmt werden.
Die Art und Weise, wie die Genauigkeit eingehalten werden soll, wird über den Parameter tm (für ' truncation mode ') im Pseudocode übergeben, die Genauigkeit selbst mit ϵ. Die Matrizen, die den Basiswechsel zur komprimierten Clusterbasis beschreiben, werden im Array C gespeichert.

procedure truncate($\mathcal{T}_\mathcal{I}, \mathcal{R}, t, V, V^{new}, Z, C, \epsilon, tm$)
  Matrix $\widetilde{V}, X, Q, O, \Sigma, \widetilde{U}$, int k^{new}, r
  for all $t' \in \operatorname{kind}(t)$ **do**
    truncate($\mathcal{T}_\mathcal{I}, \mathcal{R}, t', V, V^{new}, Z, C, \epsilon, tm$)
  end for
  for all directions $c \in \mathcal{R}_t$ **do**
    if $\operatorname{kind}(t) = \emptyset$ **then**
      $r \leftarrow \#\mathfrak{I}t, \qquad \mathbb{C}^{r \times k_{tc}} \ni \widetilde{V} \leftarrow V_{tc}|_{\star(t \times k_{tc})}$
    else $\operatorname{kind}(t) = \{t'_1, \ldots, t'_\tau\}$ with direction $c' = \mathrm{r}_t(c)$

$$r \leftarrow \sum_{i=1}^{\tau} k^{new}_{t'_i c'}, \qquad \mathbb{C}^{r \times k_{tc}} \ni \widetilde{V} \leftarrow \begin{pmatrix} C_{t'_1 c'} E_{t'_1 c} \\ \vdots \\ C_{t'_\tau c'} E_{t'_\tau c} \end{pmatrix}$$

    end if
    $\mathbb{C}^{r \times \widetilde{m}} \ni X \leftarrow \widetilde{V} Z^*_{tc}$, compute SVD of X with $X = Q\Sigma O^*$ and q singular values with minimal rank $k^{new} \in \underline{q}_\rfloor$ for tm and ϵ $\quad \triangleright\ \widetilde{m} \widehat{=}$ rows Z_{tc}
    $\widetilde{U} \leftarrow Q|_{\star(r \times k^{new})}, \qquad \mathbb{C}^{k^{new} \times k_{tc}} \ni C_{tc} = \widetilde{U}^* \widetilde{V}, \qquad k^{new}_{tc} \leftarrow k^{new}$
    if $\operatorname{kind}(t) = \emptyset$ **then**
      $V^{new}_{tc} \in \mathbb{C}^{\mathcal{I} \times k^{new}_{tc}}, \qquad V^{new}_{tc}|_{\star(t \times k^{new}_{tc})} \leftarrow \widetilde{U}$
    else
      $r = 0, \qquad$ and $c' = \mathrm{r}_t(c)$
      for all $t' \in \operatorname{kind}(t)$ **do**
        $E^{new}_{t'c} \leftarrow \widetilde{U}|_{[r+1, r+k^{new}_{t'c'}] \times k^{new}_{tc}}, \qquad r \leftarrow r + k^{new}_{t'c'}$
      end for
    end if
  end for
end procedure

Algorithmus 5.6: Kürzen der Clusterbasis

Lemma 5.17 (Aufwand des Kürzens)
Die Anzahl der nötigen Operationen beim Kürzen einer richtungsabhängigen Clusterbasis $\{V_{tc}\}_{\substack{t\in\mathcal{T}_\mathcal{I}\\ c\in\mathcal{R}_t}}$ *einer* $\mathcal{RH}^2$*-Matrix ist durch*

$$k^3\mathcal{C}_{tr}(\#\mathcal{T}_\mathcal{I} + \kappa^2(p_\mathcal{I}+1)\mathcal{C}_{tk}) \tag{5.3.4}$$

mit

$$\mathcal{C}_{tr} := 2\mathcal{C}_{kk} + \mathcal{C}_{svd}\mathcal{C}_{sr} + 4\mathcal{C}_{sr},$$

wobei $\mathcal{C}_{sr} = \max\{\mathcal{C}_{bk}, \mathcal{C}_{kk}\}$ *ist, beschränkt.*

Beweis: Seien $t \in \mathcal{T}_\mathcal{I}$ und $c \in \mathcal{R}_t$ gegeben. Falls $\operatorname{kind}(t) = \emptyset$ gilt, wird direkt mit V_{tc} weiter gerechnet, im Algorithmus gilt entsprechend $r = \#\mathfrak{I}t$, was mit 3.1.13 zu $r \leq \mathcal{C}_{bk}k$ wird.
Falls $\operatorname{kind}(t) \neq \emptyset$ gilt, wird für jedes Kind $t' \in \operatorname{kind}(t)$ das Produkt der Basiswechselmatrix $C_{t'c'}$ und der Transfermatrix $E_{t'c}$ berechnet. Der Aufwand für alle Matrix-Multiplikationen zusammen beträgt

$$\sum_{t'\in\operatorname{kind}(t)} 2k^{new}_{t'c'}k_{t'c'}k_{tc} \leq 2k^3 \sum_{t'\in\operatorname{kind}(t)} 1 \leq 2k^3\mathcal{C}_{kk}$$

und für r gilt $r = \sum_{t'\in\operatorname{kind}(t)} k^{new}_{t'c'} \leq k\mathcal{C}_{kk}$. Vor der Singulärwertzerlegung ist zur Berechnung von A^c_{tc} noch eine Matrix-Multiplikation mit Aufwand von $2\widetilde{m}rk_{tc}$ nötig. Die Singulärwertzerlegung hat einen Aufwand von $\mathcal{C}_{svd}r\widetilde{m}\min\{r,\widetilde{m}\}$. Das anschließende Bestimmen der Matrix, welche den Basiswechsel beschreibt, ist durch $2k^{new}_{tc}rk_{tc} \leq 2k^2r$ beschränkt. Es ergibt sich ein Aufwand von

$$2k^3\mathcal{C}_{kk} + \mathcal{C}_{svd}r\widetilde{m}\min\{r,\widetilde{m}\} + 2r\widetilde{m}k + 2k^2r.$$

Setze $\mathcal{C}_{sr} = \max\{\mathcal{C}_{bk}, \mathcal{C}_{kk}\}$, dann kann mit $r \leq \max\{\mathcal{C}_{bk}, \mathcal{C}_{kk}\}k = \mathcal{C}_{sr}k$ und $\widetilde{m} \leq k$ weiter abgeschätzt werden

$$\begin{aligned}
&2k^3\mathcal{C}_{kk} + \mathcal{C}_{svd}r\widetilde{m}^2 + 2r\widetilde{m}k + 2k^3\mathcal{C}_{sr}\\
&\leq 2k^3\mathcal{C}_{kk} + \mathcal{C}_{svd}\mathcal{C}_{sr}k^3 + 4\mathcal{C}_{sr}k^3\\
&= k^3\left(2\mathcal{C}_{kk} + \mathcal{C}_{svd}\mathcal{C}_{sr} + 4\mathcal{C}_{sr}\right)\\
&= k^3\mathcal{C}_{tr}.
\end{aligned}$$

Für den Gesamtaufwand betrachte jedes $t \in \mathcal{T}_\mathcal{I}$ und jede Richtung $c \in \mathcal{R}^{eff}_t$, mit Lemma 3.14 ergibt sich folgendes Resultat

$$\sum_{t\in\mathcal{T}_\mathcal{I}}\sum_{c\in\mathcal{R}^{eff}_t} k^3\mathcal{C}_{tr} \leq k^3\mathcal{C}_{tr}\sum_{t\in\mathcal{T}_\mathcal{I}} \#\mathcal{R}^{eff}_t \leq k^3\mathcal{C}_{tr}(\#\mathcal{T}_\mathcal{I} + \kappa^2(p_\mathcal{I}+1)\mathcal{C}_{tk}).$$

□

Mit Korollar 3.17 befindet sich der Aufwand in $\mathcal{O}(k^2 \#\mathcal{I} + k^3 \kappa^2 \log_2(\#\mathcal{I}))$.

Zu guter Letzt gilt es, die neuen Kopplungsmatrizen auszurechnen und damit die reduzierte $\mathcal{RH}^2$-Matrix-Darstellung zu vollenden. Dies ist schnell durch die Multiplikation mit den Matrizen C_{tc}, C_{sc}, die den Basiswechsel zur neuen Clusterbasis beschreiben, erledigt. Das Nahfeld bleibt unangetastet und muss daher auch nicht berücksichtigt werden.
Der Aufwand der Projektion zur Bestimmung der neuen Kopplungsmatrizen beschränkt sich auf zwei Matrix-Multiplikationen pro Kopplungsmatrix. Entsprechend wird für jedes $b = (t, s, c) \in \mathcal{L}^+_{\mathcal{I}\times\mathcal{I}}$

$$S_b^{new} = C_{tc} S_b C_{sc}^*$$

berechnet. Die einzelnen Multiplikationen der Basiswechselmatrizen C und der Kopplungsmatrix lassen sich durch einen Aufwand von $2k^3$ beschränken. Der Aufwand aller notwendigen Multiplikationen zusammen beträgt damit höchstens $4k^3$ Operationen. Um den Aufwand leichter abschätzen zu können, führe die eine Summe über die Blöcke $b \in \mathcal{L}^+_{\mathcal{I}\times\mathcal{I}}$ auf mehrere Summen in Clusterbäumen zurück

$$\sum_{t\in\mathcal{T}_\mathcal{I}} \sum_{s\in\mathrm{row}^+(t)} 4k^3 \leq 4k^3 \sum_{t\in\mathcal{T}_\mathcal{I}} \sum_{s\in\mathrm{row}(t)} 1.$$

Dies kann mit Hilfe von Lemma 3.15 durch

$$4k^3(\mathcal{C}_{sk}\left(\#\mathcal{T}_\mathcal{I} + p_\mathcal{I}\kappa^2 \mathcal{C}_{kk}\mathcal{C}_{Ck}\mathcal{C}_{uk}\mathcal{C}_{mk}^2|\Gamma|\right) + 1)$$

beschränkt werden.

Lemma 5.18 (Aufwand Projektion)
Der Aufwand der Berechnung der neuen Kopplungsmatrizen ist durch

$$4k^3(\mathcal{C}_{sk}\left(\#\mathcal{T}_\mathcal{I} + p_\mathcal{I}\kappa^2 \mathcal{C}_{kk}\mathcal{C}_{Ck}\mathcal{C}_{uk}\mathcal{C}_{mk}^2|\Gamma|\right) + 1) \tag{5.3.5}$$

beschränkt.

Beweis: Siehe oben.

Mit derselben Annahme wie zuvor kann dies ebenfalls in $\mathcal{O}(\#\mathcal{I}k^2 + \kappa^2 k^3 \log_2(\#\mathcal{I}))$ eingeordnet werden.

Der Ablauf der vollständigen Rekompression ist im Algorithmus 5.8 zu finden und der Gesamtaufwand der Rekompression ist im folgenden Theorem angegeben.

Theorem 5.19 (Aufwand Rekompression)

procedure project($\mathcal{T}_{\mathcal{I}\times\mathcal{I}}, \mathcal{R}, S, S^{new}, C_{row}, C_{col}$)
  for $b = (t, s, c) \in \mathcal{L}^+_{\mathcal{I}\times\mathcal{I}}$ **do**
    $\mathbb{C}^{k^{new}_{tc}\times k^{new}_{sc}} \ni S^{new}_b \leftarrow C_{tc} S_b C^*_{sc}$
  end for
end procedure

Algorithmus 5.7: Projektion der Kopplungsmatrizen

procedure recompress($\mathcal{T}_{\mathcal{I}\times\mathcal{I}}, \mathcal{R}, S, S^{new}, V, V^{new}, W, W^{new}, \epsilon, tm$)
  Sets of matrices $R_{row}, R_{col}, Z_{row}, Z_{col}, C_{row}, C_{col}$
  ortho_dclusterbasis($\mathcal{T}_{\mathcal{I}}, \mathcal{R}, \text{wurzel}(\mathcal{T}_{\mathcal{I}}), V, R_{row}$) ▷ 5.1
  ortho_dclusterbasis($\mathcal{T}_{\mathcal{I}}, \mathcal{R}, \text{wurzel}(\mathcal{T}_{\mathcal{I}}), W, R_{col}$)
  row_weights($\mathcal{T}_{\mathcal{I}\times\mathcal{I}}, \mathcal{R}, \text{wurzel}(\mathcal{T}_{\mathcal{I}}), V, Z_{row}, R_{col}, S$) ▷ 5.5
  col_weights($\mathcal{T}_{\mathcal{I}\times\mathcal{I}}, \mathcal{R}, \text{wurzel}(\mathcal{T}_{\mathcal{I}}), W, Z_{col}, R_{row}, S$)
  truncate($\mathcal{T}_{\mathcal{I}}, \mathcal{R}, \text{wurzel}(\mathcal{T}_{\mathcal{I}}), V, V^{new}, Z_{row}, C_{row}, \epsilon, tm$) ▷ 5.6
  truncate($\mathcal{T}_{\mathcal{I}}, \mathcal{R}, \text{wurzel}(\mathcal{T}_{\mathcal{I}}), W, W^{new}, Z_{col}, C_{col}, \epsilon, tm$)
  project($\mathcal{T}_{\mathcal{I}\times\mathcal{I}}, \mathcal{R}, S, S^{new}, C_{row}, C_{col}$) ▷ 5.7
end procedure

Algorithmus 5.8: Rekompression einer $\mathcal{RH}^2$-Matrix

Der Aufwand der Rekompression einer $\mathcal{RH}^2$-Matrix ist durch

$$\begin{aligned}
&2k^2 \#\mathcal{I}\,(2\mathcal{C}_{bk}(2\mathcal{C}_{sk} + \mathcal{C}_{tr} + (\mathcal{C}_{sk} + \mathcal{C}_{kk})\mathcal{C}_{wei}) + \mathcal{C}_{or}) \\
&+ 2k^3\kappa^2(p_{\mathcal{I}} + 1)\,(2\mathcal{C}_w\mathcal{C}_{sk} + \mathcal{C}_w\mathcal{C}_{wei}(\mathcal{C}_{sk} + \mathcal{C}_{kk}) + \mathcal{C}_{tk}\mathcal{C}_{tr} + \mathcal{C}_{or}) \\
&+ 2k^3(4 + \mathcal{C}_{wei})
\end{aligned}$$

beschränkt.

Beweis: Die Abschätzung folgt direkt aus der Kombination von zweimal (5.1.5), zweimal (5.3.3), zweimal (5.3.4) und einmal (5.3.5), wobei $\mathcal{T}_{\mathcal{I}}$ mit Korollar 3.17 sowie $p_{\mathcal{I}}$ durch $p_{\mathcal{I}} + 1$ beschränkt werden und für den Aufwand von (5.3.5) wird ebenfalls $\mathcal{C}_w$ genutzt. □

Alle Teilalgorithmen lassen sich durch $\mathcal{O}(k^2(\#\mathcal{I} + k\kappa^2 \log_2(\#\mathcal{I})))$ beschränken, so dass auch der gesamte Rekompressionsalgorithmus diese Komplexität teilt.

Bemerkung 33 : *(On-the-fly-Algorithmus) Ein genauer Blick auf den Algorithmus zeigt, dass alle Kopplungs-, Blatt- und Transfermatrizen nur zweimal während des Rekompressionsalgorithmus benötigt werden, so dass es lohnenswert sein kann, diese nicht zu speichern, sondern nur dann zu berechnen, wenn sie wirklich benötigt werden. Zwar wird auf diese Weise zweimal die Zeit des Aufstellens der $\mathcal{RH}^2$-Matrix benötigt, welche jedoch für den richtungsabhängigen Interpolationsansatz sehr gering ist, dafür der notwendige Speicher*

aber drastisch reduziert. Dies ermöglicht auch große Modelle oder hohe Auflösungen zu rechnen.

5.4 Numerische Experimente

In diesem Abschnitt sollen Experimente verdeutlichen, welche Auswirkungen die Komprimierungsansätze auf die Speicheranforderungen und Genauigkeit der $\mathcal{RH}^2$-Matrizen haben. Alle folgenden Rechnungen wurden auf einem *Shared Memory* System mit zwei Intel® Xeon® Platinum 8160 Prozessoren mit insgesamt 48 Kernen durchgeführt.

5.4.1 Orthogonalisierung

Bei der Orthogonalisierung der Clusterbasen handelt es sich formal nur um einen Transfer in eine effizientere Darstellung der Clusterbasen, entsprechend ist zu erwarten, dass teilweise der Rang reduziert wird, während die Genauigkeit der Approximation unangetastet bleibt. Betrachtet wird die Einheitssphäre mit unterschiedlichen Interpolationsordnungen m bei der festen Problemgröße ($n := \#\mathcal{I} = 32768$).

In Tabelle 5.1 befinden sich die Startparameter in den ersten Spalten, die Interpolationsordnung m in Spalte eins, gefolgt von der verwendeten Wellenzahl κ und dem anfänglichen Rang k. In den folgenden beiden Spalten sind der ursprüngliche Speicherbedarf für die Matrix $\widetilde{A}_e$ sowie der Speicherbedarf für die isometrische Variante $\widetilde{A}_e^o$ zu finden. Die Spalten sechs und sieben geben die benötigte Zeit zum Orthogonalisieren an, zunächst ohne Parallelisierung und dann parallelisiert. Anschließend folgt die Norm der Differenz der beiden Matrixapproximationen zur Kontrolle.
Die Abbildung 5.5 zeigt noch einmal den Speicherbedarf pro Freiheitsgrad in [KiB]/n der Standardapproximation via Interpolation gegenüber der orthogonalisierten Variante. Es ist ein eindeutig asymptotisches Verhalten mit Annäherung an den Speicherbedarf der vollbesetzten Matrix, welcher in blau eingezeichnet ist, nach der Orthogonalisierung zu erkennen. Dies legt nahe, dass noch nach weiteren Reduktionsmöglichkeiten für den Speicherbedarf zu suchen ist. Zudem zeigt sich an dieser Stelle erneut das Problem, dass die Interpolation als Approximationsansatz bei gegebener Genauigkeit schnell zu gravierend zu hohen Rängen führt. Glücklicherweise kann der siebten Spalte der Tabelle 5.1 entnommen werden, dass sich die Orthogonalisierung einer Clusterbasis gut parallelisieren lässt, so dass der Speicherbedarf mit diesem einfachen Hilfsmittel schnell, ohne den Approximationsfehler zu ändern, reduziert werden kann.
Das Experiment mit dem Einfachschichtoperator wurde noch einmal für hohe Wellenzahlen

m	κ	k	$\widetilde{A}_e$ [GiB]	$\widetilde{A}_e^o$ [GiB]	Zeit [s]	Zeit$_p$ [s]	$\|\widetilde{A}_e - \widetilde{A}_e^o\|_2$
0	8	1	1.18	1.18	0.6	0.8	1.1_{-20}
1	8	8	1.47	1.47	1.5	0.9	2.7_{-20}
2	8	27	4.54	4.16	19.9	3.5	4.4_{-20}
3	8	64	20.04	9.73	127.5	14.0	5.1_{-20}
4	8	125	73.02	14.13	484.3	28.0	5.3_{-20}
5	8	216	215.57	16.04	1316.1	56.6	5.4_{-20}

Tabelle 5.1: Auswirkung der Orthogonalisierung bei variierender Interpolationsordnung im niedrigfrequenten Fall auf der Sphäre ($n = 32768$)

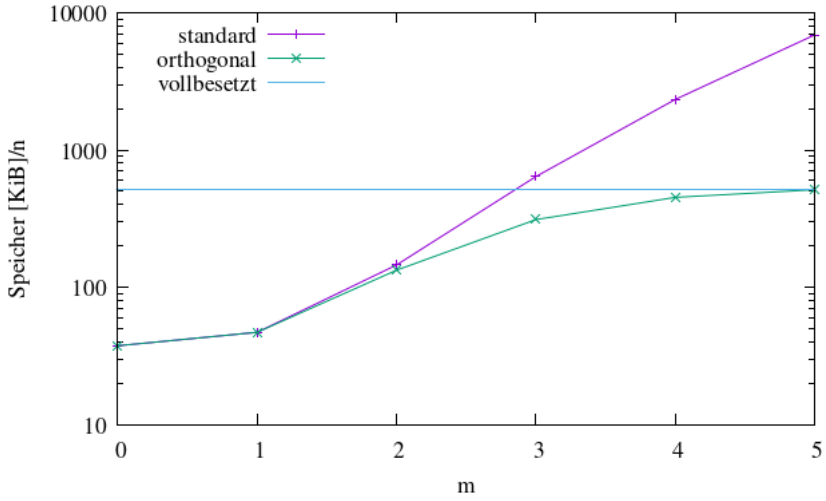


Abbildung 5.5: Vergleich der Standardapproximation und einer mit isometrischen Clusterbasen des Einfachschichtoperators im niedrigfrequenten Bereich

m	κ	k	$\widetilde{A}_e$ [GiB]	$\widetilde{A}_e^o$ [GiB]	Zeit [s]	Zeit$_p$ [s]	$\|\tilde{A}_e - \tilde{A}_e^o\|_2$
0	32	1	8.74	8.74	1.9	0.8	2.1_{-21}
1	32	8	9.38	9.38	3.8	1.1	4.8_{-21}
2	32	27	15.62	13.85	35.5	12.2	7.2_{-21}
3	32	64	46.78	16.42	163.6	20.8	7.5_{-21}
4	32	125	152.91	16.60	528.0	39.8	8.9_{-21}
5	32	216	437.98	16.60	1418.3	47.9	1.1_{-20}

Tabelle 5.2: Auswirkung der Orthogonalisierung bei variierender Interpolationsordnung im hochfrequenten Fall auf der Sphäre ($n = 32768$)

m	κ	k	$\widetilde{A}_d$ [GiB]	$\widetilde{A}_d^o$ [GiB]	Zeit [s]	Zeit$_p$ [s]	$\|\tilde{A}_d - \tilde{A}_d^o\|_2$
0	16	1	3.32	3.32	1.2	0.8	1.9_{-20}
1	16	8	3.91	3.91	2.8	0.9	9.3_{-19}
2	16	27	10.10	8.96	36.3	8.1	2.0_{-18}
3	16	64	41.36	14.61	203.9	19.9	3.6_{-18}
4	16	125	148.32	16.25	670.1	33.8	7.3_{-18}
5	16	216	436.13	16.44	1800	74.7	1.3_{-17}

Tabelle 5.3: Auswirkung der Orthogonalisierung bei variierender Interpolationsordnung für den Doppelschichtoperator auf der Sphäre ($n = 32768$)

wiederholt, die Ergebnisse sind in Tabelle 5.2 zu finden. Obwohl sich der Speicherbedarf aufgrund der schärferen parabolischen Zulässigkeitsbedingung für die reine Interpolation noch einmal erhöht, liefert die Orthogonalisierung Ergebnisse in derselben Größenordnung wie bei kleineren Wellenzahlen. Der Speicherbedarf kann um mehr als den Faktor 10 reduziert werden.
Entsprechend fällt der Unterschied der Kurven in der Abbildung 5.6 noch deutlicher aus als in Abbildung 5.5.
Eine Untersuchung des Doppelschichtoperators zeigt, dass der Algorithmus ebenso gut für eine Matrix mit unterschiedlichen Clusterbasen funktioniert. Die Ergebnisse des Experiments sind in der Tabelle 5.3 zu finden. Auch hier ist das asymptotische Verhalten im Speicher gut zu erkennen.
Eine grafische Auswertung der Tabelle 5.3 findet sich in der Abbildung 5.7.

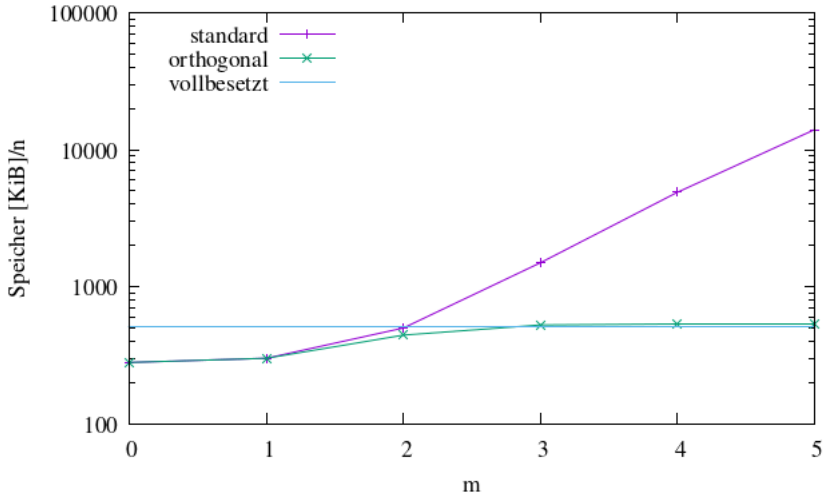


Abbildung 5.6: Vergleich der Standardapproximation und einer mit isometrischen Clusterbasen des Einfachschichtoperators im hochfrequenten Bereich

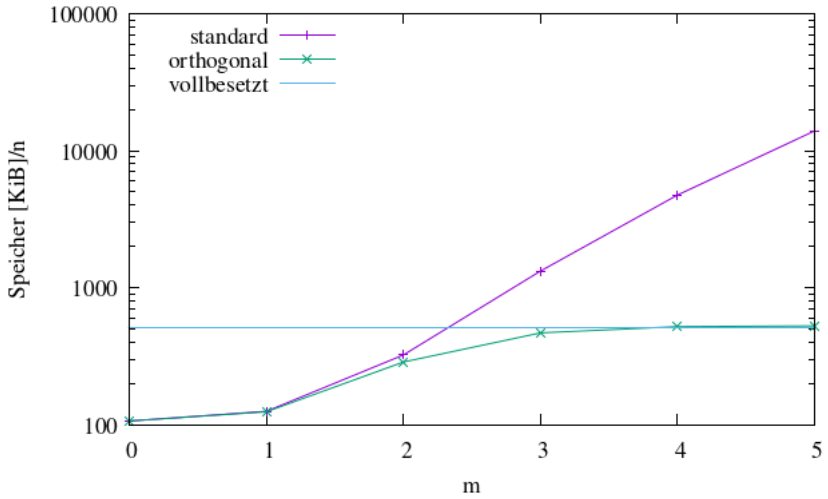


Abbildung 5.7: Vergleich der Standardapproximation und einer mit isometrischen Clusterbasen für den Doppelschichtoperator

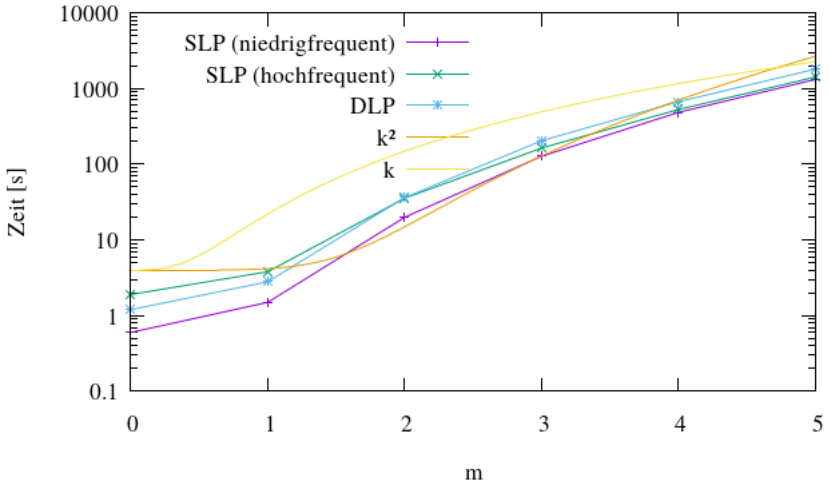


Abbildung 5.8: Vergleich der benötigten Zeit zum Orthogonalisieren

Die Abschätzung des Orthogonalisierungsalgorithmus (5.1.5) führt zu einem Aufwand von $\mathcal{O}(nk^2 + k^3\kappa^2 \log_2(n))$, der nach den durchgeführten Experimenten zu pessimistisch zu sein scheint. In der Abbildung 5.8 befinden sich die Daten des Zeitaufwands ohne Parallelisierung zu einem festen n und neben den einzelnen Zeiten noch eine Vergleichskurve aus $\mathcal{O}(k)$, die den asymptotischen Verlauf gut wiederzugeben scheint, und eine Vergleichskurve aus $\mathcal{O}(k^2)$. Wahrscheinlich ist die Abschätzung mit dem maximalen Rang zu grob und eine Variante mit gewichteten Rängen wäre schärfer, da sie dem seltenen Auftreten hoher Ränge besser gerecht wird. Auch scheint der Einfluss der Freiheitsgrade in diesem Experiment nicht so hoch zu sein, dies kann jedoch auch mit der verwendeten Geometrie oder der restriktiven parabolischen Zulässigkeitsbedingung zusammenhängen.

5.4.2 Kompression und Rekompression

Bei den folgenden Experimenten geht es darum, die Eigenschaften der Rekompression insbesondere bezüglich der Laufzeit und der Speicheranforderungen genauer zu untersuchen und den Vergleich zu den bisherigen Ansätzen zu ziehen.
Um die direkte Kompression, die Interpolation und die Rekompression miteinander zu vergleichen, werden verschiedene Wellenzahlen und Problemgrößen betrachtet. Zunächst werden einige Ergebnisse für den Einfachschichtoperator gezeigt. Dabei bezeichne mit A_e die vollbesetzte Matrix, mit $\widetilde{A}_e^c$ die direkt komprimierte, mit $\widetilde{A}_e^i$ die aus der Interpolation stammende Approximation, mit $\widetilde{A}_e^o$ die Approximation mit isometrischen Clusterbasen und mit $\widetilde{A}_e^r$ die rekomprimierte Approximation. Für die Rekompression wurde eine 'on-the-fly'-

n	κ	A_e [KiB]/n	$\widetilde{A}_e^c$ [KiB]/n	$\widetilde{A}_e^i$ [KiB]/n	$\widetilde{A}_e^o$ [KiB]/n	$\widetilde{A}_e^r$ [KiB]/n
32768	8	512.0	57.3	2336.7	451.9	56.2
32768	16	512.0	164.9	4746.1	520.0	165.2
32768	32	512.0	309.5	4893.1	531.3	309.3
73728	12	1152.0	65.3	3082.2	832.5	65.8
73728	24	1152.0	181.7	6911.0	1133.9	182.0
73728	48	1152.0	440.6	11983.0	1183.7	440.7
131072	16	2048.0	58.5	3497.1	1186.5	58.5
131072	32	2048.0	162.1	8674.2	1767.9	162.1
131072	64	2048.0	461.7	-	-	461.8

Tabelle 5.4: Vergleich des Speicheraufwands der Interpolation (m = 4) mit direkter Kompression und Rekompression

Variante des Algorithmus verwendet.

Bei dem Experiment wurde darauf geachtet, dass die Fehler der Approximation gemessen in der relativen Frobeniusnorm in derselben Größenordnung liegen. Die einzelnen Ergebnisse bezüglich des Speichers, der benötigten Zeit und des maximalen Rangs sind dabei in den Tabellen 5.4, 5.5 sowie 5.6 zusammengefasst.

Die Tabelle 5.4 zeigt die Ergebnisse für den Speicheraufwand. Die erste Spalte enthält die Freiheitsgrade, die zweite die verwendete Wellenzahl und die Folgenden den jeweiligen Speicherbedarf der Matrix beziehungsweise ihrer Approximation. In der letzten Zeile konnte der Speicher für die Interpolation und die orthogonalisierte Approximation nicht angegeben werden, da die Approximation via Interpolation nicht mehr in den Speicher der Maschine gepasst hat.
Wie zu erwarten, sind der Speicheraufwand für die direkte Kompression und der Speicheraufwand der Rekompression quasi identisch, die Rekompression liefert folglich ebenfalls gute Kompressionsraten zu einer gegebenen Genauigkeit. Weiterhin bestätigt sich noch einmal, dass die Interpolation allein deutlich zu hohe Speicheranforderungen hat und die Orthogonalisierung einen ersten Schritt in die richtige Richtung darstellt.
Ebenso von Interesse ist die Frage nach der benötigten Zeit zum Aufstellen. Die Ergebnisse für die Zeit von demselben Experiment finden sich in Tabelle 5.5.

Bei den Ergebnissen für die Zeit zeigt sich der große Vorteil der Interpolation, das Aufstellen der Matrix ist sehr schnell erledigt. Die Rekompression ist die zeitintensivste Variante, ihr

n	κ	A_e [s]	$\widetilde{A}_e^c$ [s]	$\widetilde{A}_e^i$ [s]	$\widetilde{A}_e^o$ [s]	$\widetilde{A}_e^r$ [s]
32768	8	29.2	82.3	21.5	44.2	91.5
32768	16	29.5	79.6	37.2	70.1	146.0
32768	32	29.1	53.0	43.8	76.4	123.3
73728	12	143.5	430.6	59.8	133.2	313.4
73728	24	144.8	372.8	108.3	242.6	498.8
73728	48	145.0	264.5	167.2	346.9	630.0
131072	16	449.6	1197.2	96.9	264.3	684.9
131072	32	451.0	1101.5	190.1	499.6	1134.9
131072	64	454.0	836.5	-	-	1829.9

Tabelle 5.5: Vergleich der Zeit zum Aufstellen der Interpolation (m = 4) mit direkter Kompression und Rekompression

Vorteil liegt darin, dass es sich um eine *on-the-fly*-Variante handelt und sie demnach einen geringen Speicheraufwand während des Aufstellens hat. Die hier verwendete direkte Kompression ist auf eine vorher berechnete vollbesetzte Matrix angewiesen, entsprechend setzt sich der zeitliche Aufwand der direkten Kompression aus dem Aufstellen der vollbesetzten und der Kompression zusammen und wird für eine steigende Zahl an Freiheitsgraden zunehmend unattraktiver. Weiterhin ist davon auszugehen, dass eine *on-the-fly*-Variante einen höheren Zeitaufwand hat, da teilweise zeitliche Vorteile der parallelen Berechnung der vollbesetzten Matrix verloren gehen und je nach Umsetzung eventuell mehrfaches Aufstellen von Teilmatrizen nötig wird. Auch bei dem Aufwand der Orthogonalisierung ist die Zeit, um eine Approximation via Interpolation zu berechnen, mit enthalten.
Ebenso zeigt sich noch einmal der Einfluss der Wellenzahl, der über die parabolische Zulässigkeitsbedingung (2.3.2b) stark auf den richtungsabhängigen Blockbaum wirkt. Blöcke werden erst auf höheren Stufen zulässig, so dass die zulässigen Blöcke kleiner sind und die Gesamtzahl der Blöcke steigt sowie mehr Richtungen auf den höheren Stufen des Blockbaums auftreten. Dies führt dazu, dass die Clusterbasen auf höheren Stufen mehr Richtungen zur Verfügung haben, so dass sich die Laufzeit der Orthogonalisierung und Rekompression verschlechtern.

Das gleiche Experiment wurde noch einmal für den Doppelschichtoperator wiederholt, die Ergebnisse finden sich in den Tabellen 5.7, 5.8 und 5.9.
Der benötigte Speicheraufwand ist beim Doppelschichtoperator nahezu identisch, entsprechend ist die Entwicklung des Speicheraufwands bei steigender Zahl der Freiheitsgrade

n	κ	$\frac{\|A_e-\tilde{A}_e^c\|_F}{\|A_e\|_F}$	k_{max}	$\frac{\|A_e-\tilde{A}_e^i\|_F}{\|A_e\|_F}$	k_{max}	$\frac{\|A_e-\tilde{A}_e^r\|_F}{\|A_e\|_F}$	k_{max}
32768	8	4.09_{-5}	13	2.92_{-5}	125	5.03_{-5}	13
32768	16	2.85_{-5}	17	1.15_{-4}	125	1.19_{-4}	17
32768	32	2.94_{-5}	15	2.15_{-5}	125	3.65_{-5}	15
73728	12	3.72_{-5}	16	7.46_{-5}	125	8.31_{-5}	16
73728	24	3.58_{-5}	18	9.38_{-5}	125	1.00_{-4}	18
73728	48	3.34_{-5}	16	6.67_{-5}	125	7.45_{-5}	16
131072	16	3.43_{-4}	13	1.17_{-4}	125	3.63_{-4}	13
131072	32	3.66_{-4}	15	1.42_{-4}	125	4.15_{-4}	15
131072	64	3.39_{-4}	14	-	-	3.74_{-4}	13

Tabelle 5.6: Vergleich der relativen Fehler in der Frobeniusnorm und der maximal auftretenden Ränge k_{max} beim Aufstellen via Interpolation (m = 4) sowie direkter Kompression und Rekompression

n	κ	A_d [KiB]/n	$\widetilde{A}_d^c$ [KiB]/n	$\widetilde{A}_d^i$ [KiB]/n	$\widetilde{A}_d^o$ [KiB]/n	$\widetilde{A}_d^r$ [KiB]/n
32768	8	512.0	59.1	2336.7	451.9	59.1
32768	16	512.0	170.1	4746.1	520.0	169.9
32768	32	512.0	306.4	4893.1	531.3	309.5
73728	12	1152.0	52.6	3082.2	832.5	52.6
73728	24	1152.0	151.1	6911.0	1133.9	151.1
73728	48	1152.0	397.8	11983.0	1183.7	397.7
131072	16	2048.0	59.7	3497.1	1186.5	59.8
131072	32	2048.0	163.5	8674.2	1767.9	163.6
131072	64	2048.0	457.6	-	-	461.8

Tabelle 5.7: Vergleich des Speicheraufwands der Interpolation (m = 4) mit direkter Kompression und Rekompression

n	κ	A_d [s]	$\widetilde{A}_d^c$ [s]	$\widetilde{A}_d^i$ [s]	$\widetilde{A}_d^o$ [s]	$\widetilde{A}_d^r$ [s]
32768	8	30.1	90.9	21.6	45.3	92.2
32768	16	29.7	85.1	97.6	131	145.1
32768	32	30.3	53.4	45.6	81.6	116.9
73728	12	148.0	418.2	59.6	133.6	313.3
73728	24	149.0	380.6	109.7	226.5	494.1
73728	48	146.7	294.4	172.2	335.1	628.1
131072	16	471.5	1287.3	97.8	262.5	691.5
131072	32	468.1	1223.5	192.2	509.5	1141.8
131072	64	465.1	867.7	-	-	1840.4

Tabelle 5.8: Vergleich der Zeit zum Aufstellen der Interpolation (m = 4) mit direkter Kompression und Rekompression

n	κ	$\frac{\|A_d-\widetilde{A}_d^c\|_F}{\|A_d\|_F}$	k_{max}	$\frac{\|A_d-\widetilde{A}_d^i\|_F}{\|A_d\|_F}$	k_{max}	$\frac{\|A_d-\widetilde{A}_d^r\|_F}{\|A_d\|_F}$	k_{max}
32768	8	2.31_{-6}	16	6.82_{-6}	125	7.21_{-6}	16
32768	16	3.70_{-6}	18	3.76_{-5}	125	3.78_{-5}	19
32768	32	7.55_{-6}	16	7.71_{-6}	125	1.08_{-6}	16
73728	12	1.81_{-5}	12	1.35_{-5}	125	2.25_{-5}	12
73728	24	4.09_{-5}	14	2.64_{-5}	125	4.86_{-5}	14
73728	48	7.63_{-5}	12	2.29_{-5}	125	7.97_{-5}	12
131072	16	1.94_{-5}	13	2.16_{-5}	125	2.90_{-5}	13
131072	32	4.53_{-5}	15	3.14_{-5}	125	5.51_{-5}	15
131072	64	7.53_{-5}	14	-	-	9.25_{-5}	13

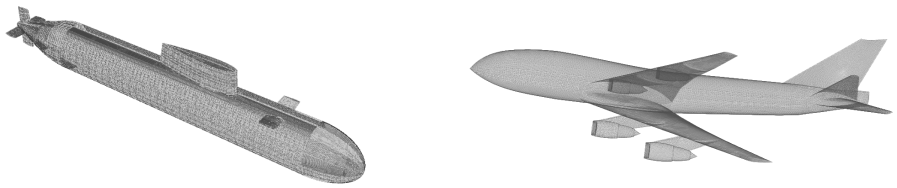
Tabelle 5.9: Vergleich der relativen Fehler in der Frobeniusnorm und der maximal auftretenden Ränge k_{max} beim Aufstellen via Interpolation (m = 4) sowie direkter Kompression und Rekompression

n	m	res	κ	k_{max}	$\widetilde{A}_e$ [GiB]	Zeit [h]	$\frac{\|A_e - \tilde{A}_e\|_F}{\|A_e\|_F}$
274920	4	16	3.15	38	29.3	1.3	5.1_{-4}
274920	5	16	3.15	46	38.7	3.2	8.9_{-5}
549836	4	32	3.15	40	53.3	1.5	1.3_{-4}

Tabelle 5.10: On-the-fly Rekompression des Einfachschichtoperators einer Boeing 747

oder wachsender Wellenzahl ebenfalls gleich. Ähnliche Ergebnisse zeigen sich auch beim zeitlichen Aufwand oder den Fehlern.

Bei großen Modellen und damit einhergehend hohen Zahlen an Freiheitsgraden ist die Rekompression in ihrem ursprünglich gedachten Einsatzgebiet. Entsprechend soll zum Schluss noch ein Ausblick auf Modelle gemacht werden, die den realen Problemen näher kommen. Dazu wurden beispielhaft Matrizen für ein Flugzeug- und ein U-Boot-Modell aufgestellt. Bilder der beiden Modelle befinden sich in Abbildung 5.9. Gerechnet wurde sowohl mit stückweise linearen als auch stückweise konstanten Basisfunktionen. Das Flugzeug (eine Boeing 747) wurde mit $\eta_1 = 10$ und $\eta_2 = 2$ sowie verschiedenen Interpolationsordnungen und Rekompressionsgenauigkeiten gerechnet. Einige Ergebnisse für die Boeing finden sich in Tabelle 5.10. Für die ersten beiden Zeilen wurden stückweise lineare Basisfunktionen gewählt, für die letzte Zeile stückweise konstante.



(a) U-Boot (b) Boeing 747

Abbildung 5.9: U-Boot und Boeing 747

Ebenso wurde mit dem U-Boot verfahren, welches mit Zulässigkeitsparametern $\eta_1 = 10$ sowie $\eta_2 = 2$ gerechnet wurde. Einige der Ergebnisse für den Fall stückweise linearer

n	m	res	κ	k_{max}	$\widetilde{A}_e$ [GiB]	Zeit [h]	$\frac{\|A_e - \tilde{A}_e\|_F}{\|A_e\|_F}$
892257	3	32	9.25	33	97.2	1.4	6.8_{-3}
892257	4	32	9.25	44	125.6	3.3	1.5_{-3}

Tabelle 5.11: On-the-fly Rekompression des Einfachschichtoperators eines U-Boots

Basisfunktion finden sich in Tabelle 5.11.

6 Vergröbern

Die bisher vorgestellten Ansätze, um den Speicheraufwand einer $\mathcal{RH}^2$-Matrix zu verringern, beruhten darauf, die einzelnen Blöcke beziehungsweise die Clusterbasen zu optimieren. Die zugrundeliegende Struktur, der richtungsabhängige Blockbaum, blieb dabei unangetastet. Es stellt sich die Frage, ob auch beim richtungsabhängigen Blockbaum Verbesserungen möglich sind, denn die parabolische Zulässigkeitsbedingung (2.3.2b) erzwingt Blockkinder, bei denen eine Referenz-$\mathcal{H}^2$-Matrix[i] schon längst zulässige Blöcke verwendet und dennoch die gewünschte Genauigkeit einhält. Dies legt den Verdacht nahe, dass die Theorie zur Zeit noch stärkere Voraussetzungen verlangt, als in den meisten praktischen Fällen nötig sind.
Entsprechend ist es Ziel der *Vergröberung*, den verwendeten Blockbaum zu überprüfen, gegebenenfalls vorhandene Blöcke zusammenzufassen und so die Gesamtzahl der Blöcke zu reduzieren.

6.1 Vergröberungsalgorithmus

Nicht alle Blöcke sind geeignet, um vergröbert zu werden. Um brauchbare Kandidaten von Blöcken für eine Vergröberung zu finden, nutze die Standardzulässigkeitsbedingung (2.3.2c) als Indiz, da die Theorie der $\mathcal{H}^2$-Matrizen nahe legt, dass zumindest diese Bedingung erfüllt sein sollte, damit eine Approximation erfolgreich sein kann.
Ist ein geeigneter Kandidat gefunden, kann die Singulärwertzerlegung genutzt werden, um zu bestimmen, ob er sich mit einem genügend kleineren Rang darstellen lässt. Schon hier zeigen sich erste Probleme dieser Idee. Handelt es sich um ein unzulässiges Blatt, kann die berechnete Singulärwertzerlegung in die nötige Matrix der Clusterbasis umgerechnet werden, entsprechend direkt aus dem unzulässigen Block ein zulässiger erstellt werden. Hat der betrachtete Block Kinder, gestaltet sich das Vorgehen deutlich schwieriger, da der Block zunächst rekonstruiert werden muss. Bei $\mathcal{RH}^2$-Matrizen reicht einfaches 'Zusammenkleben' der Teilmatrizen mit Hilfe der Kinder und der Transfermatrizen nicht mehr aus. Die Konstruktion des richtungsabhängigen Blockbaums gewährleistet nicht, dass die Kinder auch die Bestapproximation der Elternrichtung nutzen. Im schlimmsten Fall nutzen alle Kinder

[i] Die Approximation geschieht hier über eine Kompression via SVD und für einen Blockbaum, der lediglich die Standardzulässigkeitsbedingung erfüllt.

unterschiedliche Richtungen. Einzig und allein bei unzulässigen Kindern stellt die Richtung kein Problem dar, da die Richtung für vollbesetzte Matrizen irrelevant ist.
Der durch das Verwenden verschiedener Kinderrichtungen entstehende Fehler lässt sich schlecht kontrollieren, da keine scharfen Aussagen zur auftretenden Differenz der beiden Richtungen existieren. Aus diesem Grund wird nach alternativen Vorgehensweisen gesucht. Eine Möglichkeit bei zulässigen Blöcken ist es, diese zunächst via orthogonaler Projektion in die entsprechende Kinderrichtung c' umzurechnen. Nehme dazu an, dass die Clusterbasen isometrisch sind. Verwende für einen Block $b = (t, s, c)$ die projizierte Kopplungsmatrix S'_b mit

$$S'_b := V^*_{tc'} V_{tc} S_b W^*_{sc} W_{sc'},$$

so dass die Multiplikation mit den Matrizen der Clusterbasen dann die projizierte Matrix liefert

$$V_{tc'} S'_b W^*_{sc'} = V_{tc'} V^*_{tc'} V_{tc} S_b W^*_{sc} W_{sc'} W^*_{sc'} = V_{tc} S_b W^*_{sc}.$$

Leider bedeutet dieser Ansatz zusätzliche Matrix-Multiplikationen im Algorithmus und damit eine längere Laufzeit. Zudem ist unklar, in welcher Größenordnung sich der durch die Projektion auftretende zusätzliche Fehler bewegt, so dass dieser Ansatz nicht als zielführend anzusehen ist.
Deutlich praktikabler wäre es, wenn die Kinder in solchen Fällen schon die benötigte Elternrichtung hätten, also mit '*vererbten Richtungen*' gearbeitet wird und entsprechend kein Projektionsfehler auftreten würde und keine weiteren Matrix-Multiplikationen nötig wären. Dies ließe sich dadurch erreichen, dass beim Erstellen des Blockbaumes grundsätzlich bei zulässiger Standardzulässigkeitsbedingung mögliche Kinder die Richtung des Elternblocks erben. Zusätzlich wird noch eine Erweiterung der Menge der Richtungen $\mathcal{R}_t$ um die vererbten Richtungen aus niedrigeren Stufen notwendig. Praktisch bedeutet dies möglicherweise deutlich mehr Richtungen auf den höheren Stufen. Der Nachteil des zunächst auftretenden erhöhten Speichers durch eine höhere Anzahl an Richtungen sollte nach dem Vergröbern relativiert worden sein, da die Hoffnung besteht, dass viele der Kinderblöcke zugunsten der gröberen Struktur wegfallen.

Durch den veränderten Algorithmus zur Konstruktion des Blockbaums entsteht kein zeitlicher Mehraufwand. Der Zeitaufwand des Algorithmus 6.1 $build\_blocktree\_legacy$ entspricht dem des normalen Blockbaumbaus, da für eine $\mathcal{RH}^2$-Matrix generell zwei Zulässigkeitsbedingungen überprüft werden müssen. Die Berechnung der Richtung entfällt beim modifizierte Algorithmus $build\_blocktree\_constdirect$, demnach sollte hier sogar ein wenig Rechenzeit eingespart werden.
Anders gestaltet sich die Lage beim benötigten Speicher. Im schlimmsten Fall müssen die Mengen der Richtungen $\mathcal{R}_\ell$ um viele oder alle Richtungen der Stufe $\ell - 1$ erweitert werden.

function build_blocktree_legacy($t, s, \mathcal{Z}_s, \mathcal{Z}_p, \mathcal{R}$)
  Boolean z_s, z_p, block b, b', direction c_b
  $z_s \leftarrow \mathcal{Z}_s(t,s)$, $z_p \leftarrow \mathcal{Z}_p(t,s)$
  $c_b \leftarrow \hat{c} \in \mathcal{R}_t$ such that $\left\| \frac{m_t - m_s}{\|m_t - m_s\|} - \hat{c} \right\| \leq \left\| \frac{m_t - m_s}{\|m_t - m_s\|} - c \right\|$ for all $c \in \mathcal{R}_t$
  $b \leftarrow (t, s, c_b)$
  if ($z_p == false$ or $z_s == false$) and ($\mathrm{kind}(t) \neq \emptyset$ and $\mathrm{kind}(s) \neq \emptyset$) **then**
    $\mathrm{kind}(b) = \emptyset$
    for all $t' \in \mathrm{kind}(t)$ **do**
      for all $s' \in \mathrm{kind}(s)$ **do**
        if $z_s == true$ **then**
          $b' \leftarrow$ build_blocktree_constdirect($t', s', c_b, \mathcal{Z}_p, \mathcal{R}$)
        else
          $b' \leftarrow$ build_blocktree_legacy($t', s', \mathcal{Z}_s, \mathcal{Z}_p, \mathcal{R}$)
        end if
        $\mathrm{kind}(b) \leftarrow \mathrm{kind}(b) \cup b'$
      end for
    end for
  end if
  return b
end function

Algorithmus 6.1: Konstruktion eines Blockbaums mit vererbten Richtungen

```
function build_blocktree_constdirect(t, s, c, Z, R)
    Boolean z,    block b, b'
    z ← Z(t, s),    b ← (t, s, c)
    if c ∉ R_t then
        R_t ← R_t ∪ c
    end if
    if z == false and kind(t) ≠ ∅ and kind(s) ≠ ∅ then
        kind(b) = ∅
        for all t' ∈ kind(t) do
            for all s' ∈ kind(s) do
                b' ← build_blocktree_constdirect(t', s', c, Z, R)
                kind(b) ← kind(b) ∪ b'
            end for
        end for
    end if
    return b
end function
```

Algorithmus 6.2: Konstruktion eines Teilblockbaums mit konstanter Richtung

Der Aufwand kann zwar, je nach Vorgehen bei der Programmierung minimiert werden, jedoch bleibt an dieser Stelle ein erhöhter Speicheraufwand. Abhängig ist der Mehraufwand sowohl von den Zulässigkeitsbedingungen, also κ, η_1 und η_2, als auch von der Geometrie.
Es bleibt noch die Frage nach dem durch die vererbten Richtungen entstehenden Fehler. Davon ausgehend, dass die Blöcke auch wirklich vergröbert werden, also tatsächlich ein Block mit seiner optimalen Richtung entsteht, entfällt jeder Reinterpolationsfehler beim Rekonstruieren dieses Block, denn all seine Kindercluster haben in den Clusterbasen dieselbe Richtung zur Verfügung wie er. Entsprechend reduziert sich die Fehleranalyse darauf, wie gut die Kinder in der vererbten Richtung dargestellt werden.
In den Kindern selbst ist nicht klar, in wie weit die vererbte Richtung von der ansonsten gewählten abweicht. Da alle Punkte der Kinderblöcke schon im Elternblock enthalten sind, sind auch die hier auftretenden Kombinationen mit abgedeckt. Alle bisher getroffenen Fehleraussagen gelten damit weiterhin, es ändert sich nur der Parameter für die Zulässigkeitsbedingung der Richtungen. Der auftretende Fehler ist entsprechend unter Kontrolle und die exponentielle Konvergenz bleibt erhalten, selbst wenn nicht vergröbert werden sollte. Da der entstehende Fehler sich nur minimal von dem des klassischen Vorgehens unterscheidet, stellen vererbte Richtungen ein praktikables Konzept für einen Ansatz beim Vergröbern dar. Darauf aufbauend soll ein konkreter Vergröberungsalgorithmus entworfen werden.
Zunächst halte fest, dass, wenn die Anzahl der Blöcke, die mit Hilfe einer Clusterbasis dargestellt werden, wächst, potentiell mehr Informationen in den einzelnen Matrizen der

Clusterbasis enthalten sein müssen. Entsprechend muss während des Vergröberns die Möglichkeit gegeben sein, dass die Ränge wachsen können. Das Rangwachstum ist dabei so zu verstehen, dass der minimal nötige Rang anwachsen kann, nicht dass ein, zum Beispiel aus der Interpolation stammender, schon deutlich zu hoher Rang weiter wachsen soll.
Um ein Rangwachstum ermöglichen zu können, werden die vergröberten Matrizen zunächst in einer *Rang-k-Darstellung* [32, S. 26] zwischengespeichert. Rang-k-Darstellungen sind das Mittel der Wahl zur Approximation bei $\mathcal{H}$-Matrizen und ihre Arithmetik ist entsprechend erforscht und ausformuliert.

Definition 6.1 (Rang-k-Darstellung)
Falls zu einer Matrix $G \in \mathbb{C}^{\mathcal{I}\times\mathcal{J}}$ ein $k \in \mathbb{N}$ und zwei Matrizen $A \in \mathbb{C}^{\mathcal{I}\times k}, B \in \mathbb{C}^{\mathcal{J}\times k}$ existieren mit

$$G = AB^*,$$

dann bildet das Tupel (A, B) eine Rang-k-Darstellung der Matrix G.

Der Algorithmus zur Vergröberung wird grob in zwei Phasen geteilt. In der ersten werden die potentiellen Kandidaten zur Vergröberung untersucht und wenn möglich zunächst durch Rang-k-Darstellungen approximiert. Bei der zweiten Phase wird eine Abwandlung der Rekompression von $\mathcal{RH}^2$-Matrizen genutzt, um neue und optimale Clusterbasen zu finden und das ursprüngliche $\mathcal{RH}^2$-Matrixformat wieder herzustellen.
Gehe im Folgenden davon aus, dass die untersuchte $\mathcal{RH}^2$-Matrix isometrische Clusterbasen besitzt, dies ermöglicht es, den Algorithmus übersichtlicher zu gestalten. Um immer mit möglichst kleinen Matrizen arbeiten zu können, beginnt die erste Phase des Algorithmus mit der Überprüfung, ob vergröbert werden kann, in den Blättern. Ist das betrachtete Blatt schon zulässig, kann es übersprungen werden. Trifft der Algorithmus auf ein unzulässiges Blatt $b = (t, s, c)$, welches aber die Standardzulässigkeitsbedingung erfüllt, wird mit Hilfe der Singulärwertzerlegung getestet, ob es auch zur gegebenen Genauigkeit mit einer Rang-k-Darstellung approximiert werden kann. Wird eine vollbesetzte Matrix mit Hilfe der Singulärwertzerlegung auf Rang k gekürzt, wird das Ergebnis der Singulärwertzerlegung mit zwei Matrizen $A_{ts} \in \mathbb{C}^{\mathfrak{I}t\times k}$ und $B_{ts} \in \mathbb{C}^{\mathfrak{I}s\times k}$ dargestellt.

$$N_b \approx \underbrace{U\Sigma}_{=A_{ts}} \underbrace{V^*}_{=B_{ts}^*} = A_{ts}B_{ts}^*.$$

Ist eine solche Approximation möglich, werden A_{ts}, B_{ts} gespeichert, falls nicht, besteht auch keine Hoffnung, dass Vorfahrenblöcke noch vergröbert werden können, so dass diese auch nicht mehr überprüft werden müssen.
Ist der Block, der vergröbert approximiert werden soll, noch unterteilt, kann zwischen zwei Fällen unterschieden werden. Entweder seine Kinder sind selbst schon vergröbert, liegen also in Rang-k-Darstellung vor, oder mindestens eines der Kinder ist zulässig. Fälle, bei

denen noch unzulässige Kinder vorkommen, kann es nicht geben, da der Algorithmus die weitere Überprüfung dieses Teilbaums vorher abbricht.
Haben alle Kinder eine Rang-k-Darstellung, können die Kinder sukzessiv zu einer neuen Rang-k-Darstellung zusammengefasst werden. Dazu werden zwei Kinder, die sich denselben Spaltencluster teilen und mit den Darstellungen $(A_{t_1 s_1}, B_{t_1 s_1}), (A_{t_2 s_1}, B_{t_2 s_1})$ gegeben sind, miteinander zu einer Rang-$2k$-Darstellung verschmolzen

$$\begin{pmatrix} A_{t_1 s_1} B_{t_1 s_1}^* \\ A_{t_2 s_1} B_{t_2 s_1}^* \end{pmatrix} = \begin{pmatrix} A_{t_1 s_1} & 0 \\ 0 & A_{t_2 s_1} \end{pmatrix} \begin{pmatrix} B_{t_1 s_1}^* \\ B_{t_2 s_1}^* \end{pmatrix} = \widehat{A}\widehat{B}^*$$

und wenn die Kinder sich den Zeilencluster teilen, $(A_{t_1 s_1}, B_{t_1 s_1}), (A_{t_1 s_2}, B_{t_1 s_2})$,

$$\begin{pmatrix} A_{t_1 s_1} B_{t_1 s_1}^* & A_{t_1 s_2} B_{t_1 s_2}^* \end{pmatrix} = \begin{pmatrix} A_{t_1 s_1} & A_{t_1 s_2} \end{pmatrix} \begin{pmatrix} B_{t_1 s_1}^* & 0 \\ 0 & B_{t_1 s_2}^* \end{pmatrix} = \widehat{A}\widehat{B}^*.$$

Auf diese Weise können auch mehr als zwei Kinder auf einmal verschmolzen werden. Um alle Kinder zusammenzufügen, können zunächst alle Spaltenkinder, die sich einen Zeilencluster teilen, zusammengefügt werden. Anschließend können zu allen Zeilenkindern die verschmolzenen Rang-k-Darstellungen verschmolzen werden.
Der Rang der verschmolzenen Matrix $\widehat{A}\widehat{B}^*$ entspricht der Summe der einzelnen Ränge, jedoch reicht eine Approximation der Matrix $\widehat{A}\widehat{B}^*$ aus. Um den Aufwand der notwendigen Singulärwertzerlegungen zu reduzieren, kann so wie im Fall der Rekompression vorweg noch eine reduzierte QR-Zerlegung durchgeführt werden. Je nachdem, ob die Zeilen oder Spalten mehr Elemente aufweisen, wird eine QR-Zerlegung von $\widehat{A}$ oder $\widehat{B}$ bestimmt und die jeweilige andere Matrix der Rang-k-Darstellung mit der oberen Dreiecksmatrix R multipliziert. Von dieser kleineren Matrix wird die Singulärwertzerlegung berechnet und anschließend mit der isometrischen Matrix aus der QR-Zerlegung multipliziert.
Die so entstandene Rang-k-Darstellung wird mit einem weiteren Kind verschmolzen und der Vorgang des Kürzens wiederholt. Da beim Verschmelzen in jedem Schritt schon Singulärwertzerlegungen zum Kürzen genutzt werden, ist bei diesem Vorgehen keine abschließende Singulärwertzerlegung mehr notwendig. Jedoch steigt der Fehler mit jedem weiteren Schritt des Verschmelzens und Kürzens an, denn das Kürzen der Singulärwertzerlegung liefert jeweils lediglich die Bestapproximation für den betrachteten Schritt. Um die Bestapproximation für den gesamten Block zu erhalten, müssten alle Kinderblöcke in einem Schritt miteinander verschmolzen werden, was dann jedoch zu einer sehr aufwendigen Singulärwertzerlegung führen würde.
Liegt eine Mischform vor, gibt es verschiedene Möglichkeiten vorzugehen, um die Kinder auf einheitliche Form zu bringen. Da das Zielformat eine Rang-k-Darstellung ist, werden zulässige $\mathcal{RH}^2$-Matrizen in Rang-k-Darstellung umgerechnet, dazu muss nur die Kopp-

lungsmatrix an eine der Clusterbasen multipliziert werden

$$\underbrace{V_{tc}|_{\star(t\times k_{tc})}S_b}_{A_{ts}}(W_{sc}|_{\star(s\times k_{sc})})^* = A_{ts}(\underbrace{W_{sc}|_{\star(s\times k_{sc})}}_{B_{ts}})^*.$$

Dann kann erneut wie im ersten Fall sukzessiv eine Rang-k-Darstellung des Elternblocks gewonnen werden.
Sollten alle Kinder als $\mathcal{RH}^2$-Matrizen vorliegen, ist es auf Grund der vererbten Richtungen möglich und günstiger, alle Kopplungsmatrizen zusammenzufügen und von dieser Matrix die Singulärwertzerlegung zu berechnen. Da die betrachteten Clusterbasen isometrisch sind, beeinflussen die Matrizen der Clusterbasis die Singulärwerte nicht, so dass das Zusammenfügen der Kopplungsmatrizen allein ausreicht. Sollte nicht mit einer isometrischen Clusterbasis gearbeitet werden, müssen zunächst Gewichte der Clusterbasis mit Hilfe der QR-Zerlegung bestimmt werden und die jeweiligen Kopplungsmatrizen mit diesen multipliziert. Die Singulärwertzerlegung wird dann auf die Matrix der zusammengefügten gewichteten Kopplungsmatrizen angewendet. Anschließend kann die Multiplikation mit den Clusterbasen durchgeführt werden, um eine Rang-k-Darstellung zu gewinnen. Falls zunächst Gewichte der Clusterbasis bestimmt wurden, muss dies in diesem Schritt bedacht werden.
Eine grobe Übersicht des *bottom-up* Algorithmus ist in 6.3 zu finden, bei dem eine Familie von Booleans *op* verwendet wird, um zwischen potentiell vergröberbaren Blöcken und solchen, bei denen dies nicht möglich ist, zu unterscheiden.

Am Ende des Algorithmus liegt zusätzlich zum alten richtungsabhängigen Blockbaum $\mathcal{T}_{\mathcal{I}\times\mathcal{I}}$ ein neuer gekürzter Blockbaum $\mathcal{T}^{new}_{\mathcal{I}\times\mathcal{I}}$ vor. Die Blockstruktur der $\mathcal{RH}^2$-Matrix folgt dem neuen Blockbaum $\mathcal{T}^{new}_{\mathcal{I}\times\mathcal{I}}$, während die Clusterbasen noch zu dem ungekürzten Blockbaum $\mathcal{T}_{\mathcal{I}\times\mathcal{I}}$ gehören.

Lemma 6.2 (Aufwand Vergröberung eines Blocks)
Der Aufwand des Vergröberns eines Blocks nach dem oben beschriebenen Algorithmus unter der Annahme, dass nur dann vergröbert wird, wenn $k \leq \#^{\mathfrak{I}}t, \#^{\mathfrak{I}}s$ für $t, s \in \mathcal{L}_{\mathcal{I}}$ gilt und auftretende Clusterbasen isometrisch sind, ist durch

$$k^2\mathcal{C}_{vv}(\#^{\mathfrak{I}}t + \#^{\mathfrak{I}}s)$$

beschränkt, wobei die Konstante $\mathcal{C}_{vv}$

$$\mathcal{C}_{vv} := \mathcal{C}_{kk}(\mathcal{C}_{ecb} + \mathcal{C}^2_{kk}(5 + \mathcal{C}_{qr} + \mathcal{C}_{svd}))$$

erfüllt.

Beweis: Betrachte zunächst ein Blatt. Handelt es sich um ein zulässiges Blatt, ist nichts zu tun. Ist das Blatt jedoch unzulässig und ein Kandidat zum Vergröbern, wird der Algorithmus mit einer Singulärwertzerlegung prüfen, ob sich der Rang reduzieren lässt. Dazu

procedure build_coarsen($\mathcal{T}_{\mathcal{I}\times\mathcal{I}}, b, V, S, W, N, Rk, \epsilon, op$)
  Int k_{new}, matrices M, U, Σ, V
  for all $b' \in \operatorname{kind}(b)$ **do**
    build_coarsen($\mathcal{T}_{\mathcal{I}\times\mathcal{I}}, b', V, S, W, N, Rk, \epsilon, op$)
  end for
  if $op_b ==$ true **then**
    if $N_b \neq \emptyset$ **then**
      Compute SVD $N_b = U\Sigma V^*$ for given accuracy ϵ with new rank k_{new}
      if k_{new} small enough **then**
        For $Rk = (A_{ts}, B_{ts})$ do $A_{ts} \leftarrow U\Sigma$ and $B_{ts} \leftarrow V$, $b \leftarrow$ admissible and delete matrix N_b
      else
        $op_b \leftarrow$false
      end if
    else if $\operatorname{kind}(b) \neq \emptyset$ and for all children b' $op_{b'} ==$ true **then**
      Merge all children matrices together in M and compute SVD $M = U\Sigma V^*$ for given accuracy ϵ with new rank k_{new}
      if k_{new} small enough **then**
        For $Rk = (A_{ts}, B_{ts})$ do $A_{ts} \leftarrow U\Sigma$ and $B_{ts} \leftarrow V$. If all children have a coupling matrix, perform forward transformation for every child with the corresponding part of A_{ts} and B_{ts}, b gets admissible, delete all children of $b = (t', s', c)$ and their matrices $S_{b'}$ or $(A_{t's'}, B_{t's'})$
      else
        $op_b \leftarrow$false
      end if
    end if
  end if
end procedure

Algorithmus 6.3: Vergröbern

nehme ohne Beschränkung der Allgemeinheit an, dass s ein Blattcluster ist, die Singulärwertzerlegung der vollbesetzten Matrix braucht dann weniger als

$$\mathcal{C}_{svd}(\#^{\mathfrak{I}}t)(\#^{\mathfrak{I}}s)\min\left\{\#^{\mathfrak{I}}t,\#^{\mathfrak{I}}s\right\} \leq \mathcal{C}_{svd}(\#^{\mathfrak{I}}t)(\#^{\mathfrak{I}}s)^2$$

Operationen. Mit (3.1.13) kann die Anzahl der Elemente des Blattclusters s weiter abgeschätzt werden

$$\mathcal{C}_{svd}(\#^{\mathfrak{I}}t)(\#^{\mathfrak{I}}s)^2 \leq \mathcal{C}_{svd}\mathcal{C}_{bk}^2k^2(\#^{\mathfrak{I}}t).$$

Wenn eine Niedrigrang-Approximation möglich ist, wird anschließend noch die Multiplikation der Singulärwertmatrix mit der Matrizen der linken Singulärvektoren nötig, um eine Rang-k-Darstellung zu erhalten. Da es sich um eine Diagonalmatrix handelt, braucht diese Multiplikation weniger als

$$(\#^{\mathfrak{I}}t)k$$

Operationen. Insgesamt ist der Aufwand in diesem Fall durch

$$(\#^{\mathfrak{I}}t)k\left(1+k\mathcal{C}_{svd}\mathcal{C}_{bk}^2\right)$$

beschränkt.
Falls der betrachtete Block Kinder hat und zu den potentiellen Kandidaten zählt, müssen je nach Vorgehen Fallunterscheidungen gemacht werden. Wenn alle Kinder in Rang-k-Darstellung vorliegen, werden sie in Teilschritten zusammengefügt, mit einer reduzierten QR-Zerlegung verkleinert und mit Hilfe der Singulärwertzerlegung gekürzt. Die reduzierte QR-Zerlegung für das spaltenweise Zusammenfügen von zwei Kinderblöcken s_1, s_2 benötigt weniger als

$$\mathcal{C}_{qr}(k_{s_1}+k_{s_2})^2(\#^{\mathfrak{I}}s_1+\#^{\mathfrak{I}}s_2) \leq \mathcal{C}_{qr}4k^2(\#^{\mathfrak{I}}s_1+\#^{\mathfrak{I}}s_2)$$

Operationen. Werden für alle Spaltenkinder, die sich ein Zeilenkind teilen, QR-Zerlegungen bestimmt, kostet dies weniger als

$$\mathcal{C}_{qr}\mathcal{C}_{kk}^2k^2(\#^{\mathfrak{I}}s)$$

Operationen. Die Multiplikationen der A_{t_i} Matrizen der Rang-k-Darstellungen mit den jeweiligen Dreiecksmatrizen R_{t_i} aus den QR-Zerlegungen für alle Spaltenkinder und ein Zeilenkind t_i benötigt weniger als

$$2\mathcal{C}_{kk}^2k^2(\#^{\mathfrak{I}}t_i)$$

Operationen. Für die anschließende Singulärwertzerlegung werden

$$\mathcal{C}_{svd}\mathcal{C}_{kk}^2k^2(\#^{\mathfrak{I}}t_i)$$

Operationen benötigt.
Um die Verschmelzung abzuschließen und wieder eine Rang-k-Darstellung zu erhalten, müssen noch die isometrische Matrix Q^* der QR-Zerlegung von rechts mit der Matrix der rechten Singulärvektoren multipliziert werden sowie die Singulärwerte selbst an eine der beiden Matrizen der Singulärvektoren. Dies benötigt weniger als

$$k(\#^{\mathfrak{I}} t_i) + 2\mathcal{C}_{kk} k^2 (\#^{\mathfrak{I}} s)$$

Operationen. Insgesamt kann dies nach oben mit

$$\mathcal{C}_{kk}^2 k^2 \left(\#^{\mathfrak{I}} s(\mathcal{C}_{qr} + 2) + \#^{\mathfrak{I}} t_i (3 + \mathcal{C}_{svd}) \right)$$

abgeschätzt werden.
Dieser Aufwand wird für alle Zeilenkinder t_i nötig, so dass sich eine Schranke von

$$\mathcal{C}_{kk}^3 k^2 (\#^{\mathfrak{I}} s)(\mathcal{C}_{qr} + 2) + \mathcal{C}_{kk}^2 k^2 (\#^{\mathfrak{I}} t)(3 + \mathcal{C}_{svd})$$

ergibt.
Abschließend müssen noch alle Zeilen miteinander verschmolzen werden. Da hier dieselben Schritte notwendig sind, ergibt sich mit denselben Argumenten ein Aufwand von

$$\mathcal{C}_{kk}^2 k^2 \left(\#^{\mathfrak{I}} t(\mathcal{C}_{qr} + 2) + \#^{\mathfrak{I}} s(3 + \mathcal{C}_{svd}) \right).$$

Der Gesamtaufwand für diesen Fall kann damit durch

$$\begin{aligned} &\mathcal{C}_{kk}^3 k^2 (\#^{\mathfrak{I}} t + \#^{\mathfrak{I}} s)(\mathcal{C}_{qr} + 2) + \mathcal{C}_{kk}^2 k^2 (\#^{\mathfrak{I}} t + \#^{\mathfrak{I}} s)(3 + \mathcal{C}_{svd}) \\ &\quad \leq \mathcal{C}_{kk}^3 k^2 (\#^{\mathfrak{I}} t + \#^{\mathfrak{I}} s)(5 + \mathcal{C}_{qr} + \mathcal{C}_{svd}) \end{aligned}$$

beschränkt werden.
Liegen alle Kindermatrizen im $\mathcal{RH}^2$-Matrixformat vor, werden die Kopplungsmatrizen zu einer großen Matrix zusammengeklebt und von dieser Matrix die Singulärwertzerlegung berechnet. Dies benötigt weniger als

$$\mathcal{C}_{svd} \mathcal{C}_{kk}^3 k^3$$

Operationen. Anschließend müssen noch die Matrix der Clusterbasis mit der jeweilige Matrix mit Singulärvektoren multipliziert werden sowie die Singulärwerte selbst mit einer der beiden Matrizen der neuen gekürzten Rang-k-Darstellung. Dies kann mit Lemma 5.11 durch

$$k^2 (\#^{\mathfrak{I}} t) \mathcal{C}_{ecb} + k^2 (\#^{\mathfrak{I}} s) \mathcal{C}_{ecb} + k(\#^{\mathfrak{I}} t)$$

beschränkt werden. Insgesamt ergibt sich in diesem Fall

$$\mathcal{C}_{svd} \mathcal{C}_{kk}^3 k^3 + k^2 (\#^{\mathfrak{I}} t + \#^{\mathfrak{I}} s) \mathcal{C}_{ecb} + k(\#^{\mathfrak{I}} t),$$

was mit der Annahme, dass der Rang $k \leq \#^{\mathfrak{I}} t, \#^{\mathfrak{I}} s$ erfüllt, durch

$$\begin{aligned}\mathcal{C}_{svd}\mathcal{C}_{kk}^3 k^2(\#^{\mathfrak{I}} t + \#^{\mathfrak{I}} s) + k^2(\#^{\mathfrak{I}} t + \#^{\mathfrak{I}} s)(1 + \mathcal{C}_{ecb}) \\ \leq k^2(\#^{\mathfrak{I}} t + \#^{\mathfrak{I}} s)(1 + \mathcal{C}_{ecb} + \mathcal{C}_{kk}^3\mathcal{C}_{svd})\end{aligned}$$

abgeschätzt werden kann. Im letzten Fall haben die Kinder gemischte Formate. Da es relativ einfach ist, eine $\mathcal{RH}^2$-Matrix in eine Rang-k-Darstellung umzurechnen, werden alle auftretenden $\mathcal{RH}^2$-Matrixkinder umgerechnet. Das Umrechnen der Teilmatrix zum Block (t_i, s_j) kann erneut mit Lemma 5.11 abgeschätzt werden, dabei braucht die Multiplikation mit der Matrix einer der Clusterbasen und das eventuelle Rekonstruieren der Matrix der anderen Clusterbasis maximal

$$k^2(\#^{\mathfrak{I}} t_i + \#^{\mathfrak{I}} s_j)\mathcal{C}_{ecb}$$

Operationen. Im aufwendigsten denkbaren Fall müssen alle bis auf eine Teilmatrix des Blocks (t, s, c_b) zunächst in Rang-k-Darstellung umgerechnet werden, so dass alle nötigen Umrechnungen durch

$$\begin{aligned}\sum_{t' \in \mathrm{kind}(t)} \sum_{s' \in \mathrm{kind}(s)} & k^2(\#^{\mathfrak{I}} t' + \#^{\mathfrak{I}} s')\mathcal{C}_{ecb} \\ &\leq \sum_{t' \in \mathrm{kind}(t)} k^2\mathcal{C}_{ecb}(\mathcal{C}_{kk}\#^{\mathfrak{I}} t' + \#^{\mathfrak{I}} s) \\ &\leq k^2\mathcal{C}_{kk}\mathcal{C}_{ecb}(\#^{\mathfrak{I}} t + \#^{\mathfrak{I}} s)\end{aligned}$$

beschränkt werden können. Im gemischten Fall ergibt sich der Aufwand durch die Addition des Aufwands für das Umrechnen zusammen mit dem Aufwand für das Verschmelzen der Rang-k-Darstellungen

$$\begin{aligned}k^2\mathcal{C}_{kk}\mathcal{C}_{ecb}(\#^{\mathfrak{I}} t + \#^{\mathfrak{I}} s) + \mathcal{C}_{kk}^3 k^2(\#^{\mathfrak{I}} t + \#^{\mathfrak{I}} s)(5 + \mathcal{C}_{qr} + \mathcal{C}_{svd}) \\ \leq \mathcal{C}_{kk}k^2(\#^{\mathfrak{I}} t + \#^{\mathfrak{I}} s)(\mathcal{C}_{ecb} + \mathcal{C}_{kk}^2(5 + \mathcal{C}_{qr} + \mathcal{C}_{svd})).\end{aligned}$$

Dies stellt auch den maximalen Aufwand für das Vergröbern eines Blocks dar. □

Der exakte Aufwand für einen gesamten Blockbaum kann nicht genau abgeschätzt werden, da vorab nicht bekannt ist, welche Blöcke tatsächlich vergröbert werden können. Auch die Zahl der potentiellen Kandidaten kann nicht explizit angegeben werden. Bezeichne den Blockbaum, der durch die alleinige Betrachtung der Standardzulässigkeitsbedingung entsteht, mit $\mathcal{T}^s_{\mathcal{I}\times\mathcal{I}}$ und sei durch $\mathcal{L}^{+,s}_{\mathcal{I}\times\mathcal{I}}$ die Menge seiner zulässigen Blätter gegeben. Dann ist die maximale Menge, der zu überprüfenden Blöcke durch

$$\mathcal{M}_{\mathcal{I}\times\mathcal{I}} := \mathcal{T}_{\mathcal{I}\times\mathcal{I}} \setminus \left\{ \mathcal{T}^s_{\mathcal{I}\times\mathcal{I}} \setminus \left\{ \mathcal{L}^{+,s}_{\mathcal{I}\times\mathcal{I}} \right\} \right\}$$

gegeben. Entsprechend kann der Aufwand immer grob durch

$$\sum_{(t,s,c)\in\mathcal{M}_{\mathcal{I}\times\mathcal{I}}} k^2\mathcal{C}_{vv}(\#^{\mathfrak{I}}t + \#^{\mathfrak{I}}s) \tag{6.1.1}$$

beschränkt werden.

Nachdem der Aufwand des Vergröberns bestimmt ist, bleibt die Frage nach dem dabei auftretenden Fehler. Der Einsatz der Singulärwertzerlegung beim Vergröbern garantiert, dass der Fehler in jedem einzelnen Schritt kontrolliert werden kann. Wird ein unzulässiges Blatt $b = (t, s, c) \in \mathcal{T}_{\mathcal{I}}$ mit algebraischem Rang $p \in \mathbb{N}$ mit Hilfe einer aus der Singulärwertzerlegung von N_b gewonnenen Rang-k-Darstellung (A_{ts}, B_{ts}) mit $k \leq p$ approximiert, folgt für den Fehler

$$\begin{aligned}\|N_b - A_{ts}B_{ts}^*\| &= \|U\operatorname{diag}(\sigma_1,\dots,\sigma_p)O^* - U|_{\star(t\times k)}\operatorname{diag}(\sigma_1,\dots,\sigma_k)(O|_{\star(s\times k)})^*\| \\ &= \|U\operatorname{diag}(\sigma_1,\dots,\sigma_p)O^* - U\operatorname{diag}(\sigma_1,\dots,\sigma_k,0\dots,0)O^*\| \\ &= \|\operatorname{diag}(0,\dots,0,\sigma_{k+1},\dots\sigma_p)\|.\end{aligned}$$

Je nachdem ob die Frobenius- oder Spektralnorm betrachtet wird, ist der Fehler dann durch

$$\|N_b - A_{ts}B_{ts}^*\| = \begin{cases}\sigma_{k+1} & \text{für } \|\cdot\|_2 \\ \left(\sum_{i=k+1}^{p}\sigma_i^2\right)^{\frac{1}{2}} & \text{für } \|\cdot\|_F\end{cases}$$

gegeben. Auch im Fall, dass alle Kinderblöcke zulässig sind und zum Kürzen zu der Kopplungsmatrix $\widehat{S} \in \mathbb{C}^{q\times r}$ verschmolzen werden, ergibt sich direkt eine Aussage zum Fehler. Dabei seien die Kinder der Cluster t, s und die Zeilen q und Spalten r der Matrix $\widehat{S}$ durch

$$\operatorname{kind}(t) = \{t_1,\dots,t_\tau\}, \operatorname{kind}(s) = \{s_1,\dots,s_\rho\} \qquad \text{und} \qquad q = \sum_{i=1}^{\tau} k_{t_ic}, r = \sum_{i=1}^{\rho} k_{s_ic}$$

gegeben. Auch die beteiligten Matrizen der Clusterbasen werden zu $\widehat{V} \in \mathbb{C}^{\mathfrak{I}t\times q}$ und $\widehat{W} \in \mathbb{C}^{\mathfrak{I}s\times r}$ mit

$$\widehat{V} = (V_{t_1c}\cdots V_{t_\tau c})\,|_{\star(t\times q)} \qquad \text{und} \qquad \widehat{W} = \left(W_{s_1c}\cdots W_{s_\rho c}\right)|_{\star(s\times r)}$$

verschmolzen. Da isometrische Clusterbasen verwendet werden, sind auch $\widehat{V}, \widehat{W}$ isometrisch und können beim Betrachten der Norm weggelassen werden. Es reicht also aus, den Fehler der gekürzten Singulärwertzerlegung zu betrachten

$$\begin{aligned}&\|\widehat{V}\widehat{S}\widehat{W}^* - A_{ts}B_{ts}^*\| = \|\widehat{V}\widehat{S}\widehat{W}^* - \widehat{V}U|_{\star(q\times k)}\operatorname{diag}(\sigma_1,\dots,\sigma_k)(O|_{\star(r\times k)})^*\widehat{W}^*\| \\ &= \|\widehat{S} - U|_{\star(q\times k)}\operatorname{diag}(\sigma_1,\dots,\sigma_k)(O|_{\star(r\times k)})^*\| = \|\operatorname{diag}(0,\dots,0,\sigma_{k+1},\dots\sigma_p)\|.\end{aligned}$$

Erneut ergibt sich

$$\|\widehat{V}\widehat{S}\widehat{W}^* - A_{ts}B_{ts}^*\| = \begin{cases} \sigma_{k+1} & \text{für } \|\cdot\|_2, \\ \left(\sum_{i=k+1}^{p} \sigma_i^2\right)^{\frac{1}{2}} & \text{für } \|\cdot\|_F. \end{cases}$$

Wenn Rang-k-Darstellungen der Kinder eines Blocks b in einem Schritt zu einer Rang-k-Darstellung von b selbst verschmolzen und gekürzt werden, garantiert die Singulärwertzerlegung direkt eine Kontrolle des Fehlers.
Ein wenig aufwendiger ist die Analyse des Fehlers, wenn die Vergröberung über mehrere Stufen beziehungsweise beim Verschmelzen von Rang-k-Matrizen über mehrere Schritte hinweg verläuft. Um auch in diesem Fall den Fehler angeben zu können, führe die Matrix $\widehat{A}^\ell$ ein. Die Matrix $\widehat{A}^\ell$ stellt die Approximation dar, nachdem die Stufen $p_\mathcal{I}$ bis $p_\mathcal{I} - \ell$ der ursprünglichen Approximation $\widehat{A}$ vergröbert wurden. Wird der Fehler beim schrittweise Verschmelzen von Rang-k-Matrizen betrachtet, stellt $\widehat{A}^\ell$ die Approximation nach dem $(\ell+1)$-ten Schritt des Verschmelzens dar. Mit dieser Matrix analysiere die Vergröberung Schritt für Schritt. Zusätzlich nutze aus, dass sich die Spektralnorm auf einem Block $b \in \mathcal{T}_{\mathcal{I}\times\mathcal{I}} \setminus \mathcal{L}_{\mathcal{I}\times\mathcal{I}}$ durch die Normen der Teilmatrizen seiner Kinder beschränken lässt

$$\|A|_b\|_2 = \left\| \sum_{b' \in \operatorname{kind}(b)} A|_{b'} \right\|_2 \overset{\Delta}{\leq} \sum_{b' \in \operatorname{kind}(b)} \|A|_{b'}\|_2.$$

Lemma 6.3 (Fehler Vergröberung)
Sei für jeden Block $b \in \mathcal{T}_{\mathcal{I}\times\mathcal{I}}$ eine Fehlerschranke $\epsilon_b \in \mathbb{R}_{>0}$ gegeben, so dass die Vergröberung des Blocks b

$$\|\widehat{A}|_b - \widehat{A}^\ell|_b\| \leq \epsilon_b \qquad \textit{für } b \in \mathcal{L}^-_{\mathcal{I}\times\mathcal{I}} \textit{ mit } \operatorname{stufe}(b) = p_\mathcal{I} - \ell$$

beziehungsweise

$$\|\widehat{A}^\ell|_b - \widehat{A}^{\ell+1}|_b\| \leq \epsilon_b \qquad \textit{für } b \in \mathcal{T}_{\mathcal{I}\times\mathcal{I}} \setminus \mathcal{L}_{\mathcal{I}\times\mathcal{I}} \textit{ mit } \operatorname{stufe}(b) = p_\mathcal{I} - (\ell+1)$$

in der Frobenius- oder Spektralnorm erfüllt. Dann ist für jeden Block $b \in \mathcal{T}_{\mathcal{I}\times\mathcal{I}}$ der Fehler beim Vergröbern der Matrix $\widehat{A}$ in der Frobenius- oder Spektralnorm durch

$$\|\widehat{A}|_b - \widehat{A}^{q+p}|_b\| \leq \sum_{n=0}^{p} \sum_{b' \in \mathcal{T}_b^{p-n}} \epsilon_{b'}$$

beschränkt, wobei p die Baumtiefe vom Teilbaum $\mathcal{T}_b$ und $q := p_\mathcal{I} - \operatorname{stufe}(b) - p$ sei.

Beweis: Vorweg mache noch eine kleine Beobachtung. Sei ein Block $b \in \mathcal{T}_{\mathcal{I}\times\mathcal{I}}$ gegeben und sei p die Baumtiefe des Teilbaums $\mathcal{T}_b$. Falls $q := p_{\mathcal{I}} - \text{stufe}(b) - p > 0$, gilt für alle $\widehat{b} \in \mathcal{T}_{\mathcal{I}\times\mathcal{I}}$ mit $\text{stufe}(\widehat{b}) > \text{stufe}(b) + p$, dass $\widehat{b} \notin \mathcal{T}_b$. Das heißt, Vergröberungsschritte, die die q höchsten Stufen des Baums $\mathcal{T}_{\mathcal{I}\times\mathcal{I}}$ betreffen, haben keine Auswirkungen auf die Vergröberung des Blocks b, entsprechend gilt

$$\widehat{A}|_b = \widehat{A}^u|_b \qquad \text{für alle } u < q.$$

Zeige die Behauptung per abschnittsweiser Induktion über die Baumtiefe des Teilbaums $\mathcal{T}_b$.
I.A. Sei $p = 0$, dann handelt es sich bei b um ein Blatt des Teilbaum $\mathcal{T}_b$ und damit auch um ein Blatt des Baums $\mathcal{T}_{\mathcal{I}\times\mathcal{I}}$. Entsprechend gilt

$$\text{stufe}(b) = \text{stufe}(b) + p = p_{\mathcal{I}} - q.$$

Falls b ein zulässiges Blatt sein sollte, wird keine Veränderung vorgenommen, es gilt $\widehat{A}|_b = \widehat{A}^q|_b$. Wenn es sich bei b um ein unzulässiges Blatt handelt, folgt mit der ersten Annahme

$$\|\widehat{A}|_b - \widehat{A}^{q+p}|_b\| = \|\widehat{A}|_b - \widehat{A}^q|_b\| \leq \epsilon_b = \sum_{b' \in \mathcal{T}_b^{p-0}} \epsilon_{b'}.$$

I.V. Sei $p \in \underline{p_{\mathcal{I}} - 1}_{\rfloor 0}$ so gegeben, dass die die Behauptung für alle Teilbäume mit Baumtiefe p gilt.
I.S. Betrachte ein $b \in \mathcal{T}_{\mathcal{I}\times\mathcal{I}}$, dessen Teilbaum $\mathcal{T}_b$ eine Tiefe von $p+1$ hat. Nutze die Matrix des vorherigen Vergröberungsschritts $\widehat{A}^{q+p}$, um den Fehler aufzuteilen. Es gilt

$$\begin{aligned}\|\widehat{A}|_b - \widehat{A}^{q+p+1}|_b\| &= \|\widehat{A}|_b - \widehat{A}^{q+p}|_b + \widehat{A}^{q+p}|_b - \widehat{A}^{q+p+1}|_b\| \\ &\overset{\Delta}{\leq} \|\widehat{A}|_b - \widehat{A}^{q+p}|_b\| + \|\widehat{A}^{q+p}|_b - \widehat{A}^{q+p+1}|_b\|.\end{aligned}$$

Die erste Norm kann mit Hilfe der Induktionsvoraussetzung abgeschätzt werden. Zerlege den Block b dazu in seine Kinder, deren Teilbäume dann maximal eine Baumtiefe von p haben. Es folgt

$$\begin{aligned}\|\widehat{A}|_b - \widehat{A}^{q+p}|_b\| &\overset{\Delta}{\leq} \sum_{b' \in \text{kind}(b)} \|\widehat{A}|_{b'} - \widehat{A}^{q+p}|_{b'}\| \overset{I.V.}{\leq} \sum_{b' \in \text{kind}(b)} \sum_{n=0}^{p} \sum_{\hat{b} \in \mathcal{T}_{b'}^{p-n}} \epsilon_{\hat{b}} \\ &= \sum_{n=0}^{p} \sum_{b' \in \mathcal{T}_b^{(p+1)-n}} \epsilon_{b'}.\end{aligned}$$

Bei der zweiten Norm wird nur ein Schritt der Vergröberung betrachtet, welcher mit der zweiten Annahme beschränkt werden kann

$$\|\widehat{A}^{q+p}|_b - \widehat{A}^{q+p+1}|_b\| \leq \epsilon_b.$$

Insgesamt folgt

$$\|\widehat{A}|_b - \widehat{A}^{q+p+1}|_b\| \leq \sum_{n=0}^{p} \sum_{b' \in \mathcal{T}_b^{(p+1)-n}} \epsilon_{b'} + \epsilon_b = \sum_{n=0}^{p+1} \sum_{b' \in \mathcal{T}_b^{(p+1)-n}} \epsilon_{b'}.$$

□

Bemerkung 34 (Fester Fehler): *Wird in Lemma 6.3 mit einer festen Fehlerschranke $\epsilon = \epsilon_{b'}$ für alle $b' \in \mathcal{T}_b$ gearbeitet, folgt direkt*

$$\|\widehat{A}|_b - \widehat{A}^{q+p}|_b\| \leq \sum_{n=0}^{p} \sum_{b' \in \mathcal{T}_b^{p-n}} \epsilon = \#\mathcal{T}_b \epsilon,$$

wobei p die Baumtiefe von $\mathcal{T}_b$ und $q := p_{\mathcal{I}} - \operatorname{stufe}(b) - p$ sei. Entsprechend sollte bei Teilbäumen $\mathcal{T}_b$, bei denen mit vielen Vergröberungsschritten zu rechnen ist, mit angepassten, zum Beispiel stufenabhängigen Fehlerschranken $\epsilon_{b'}$ gearbeitet werden. Bei einer geschickten Wahl der Fehlerschranken $\epsilon_{b'}$ kann die Summe der Fehler auch auf eine geometrische Reihe zurückgeführt werden.

Bemerkung 35 (Fehler beim gekürzten Zusammenfügen von Rang-k-Darstellungen): *Werden mehrere Rang-k-Darstellungen der Kinder eines Blocks b nacheinander zusammengefügt, kann ebenfalls das Vorgehen aus dem Lemma 6.3 genutzt werden, um den Fehler zu kontrollieren. Wird in jedem Schritt des Zusammenfassens eine Genauigkeit ϵ_b eingehalten, ergibt sich für den Fehler nach $\ell + 1$ Schritten des Verschmelzens*

$$\|\widehat{A}|_b - \widehat{A}^{\ell}|_b\| \leq \|\widehat{A}|_b - \widehat{A}^{0}|_b\| + \sum_{n=1}^{\ell} \|\widehat{A}^{n-1}|_b - \widehat{A}^{n}|_b\| \leq (\ell+1)\epsilon_b.$$

Dabei bezeichne $\widehat{A}^n|_b$ in diesem Fall die Matrix nach dem $(n+1)$-ten Schritt des Verschmelzens.

In der zweiten Stufe des Algorithmus gilt es, die Matrix zurück in ein einheitliches $\mathcal{RH}^2$-Matrix-Format zu überführen. Dazu wird der Algorithmus der Rekompression dahingehend modifiziert, dass er auch mit Teilmatrizen in Rang-k-Darstellung arbeiten kann. Als erstes wird der Algorithmus 5.5 zur Bestimmung der Gewichte für die Rekompression angepasst. Viele Änderungen sind dabei nicht nötig, es müssen nur zusätzlich zu den Gewichten der Rekompression auch eine Art Gewichte aus den Rang-k-Darstellungen bestimmt und im Baum in die Blätter getragen werden.
Erneut nehme an, dass die Clusterbasen isometrisch sind. Ist ein Cluster t an zulässigen Blöcken mit Rang-k-Darstellung beteiligt, werden in diesen Blöcken QR-Zerlegungen von den jeweiligen B Matrizen der Rang-k-Darstellung bestimmt

$$A_{ts} B_{ts}^* = A_{ts} R_s^* Q_s^*,$$

da nur die linken Singulärvektoren von Interesse sind, kann die Matrix Q_s im Folgenden weggelassen werden. Wenn t an $\sigma \in \mathbb{N}$ Blöcken mit Rang-k-Darstellung beteiligt ist, werden alle komprimierten Blöcke zusammengeklebt

$$\widehat{A}_{tc} = \begin{pmatrix} A_{ts_1} R_{s_1}^* & A_{ts_2} R_{s_2}^* & \cdots & A_{ts_\sigma} R_{s_\sigma}^* \end{pmatrix}.$$

Da die Matrix $\widehat{A}_{tc}$ genau $\#^{\mathfrak{I}} t$ Zeilen hat, kann sie einfach durch Aufteilen an die jeweiligen Kinder weiter gegeben werden. Im jeweiligen Kind wird die Information des Elternblocks dann einfach zu der eigenen hinzugefügt. Auf diese Weise kann die Information im Baum nach unten zu den Blättern weiter getragen werden.
Wie bei der Rekompression wird für die Spaltenclusterbasis die adjungierte Matrix betrachtet. Entsprechend wird für einen Block mit Rang-k-Darstellung $A_{ts} B_{ts}^*$ die QR-Zerlegung von A_{ts} berechnet und $B_{ts} R_t^*$ für die totale richtungsabhängige Clusterbasis verwendet.

Lemma 6.4 (Aufwand Berechnung der Gewichte)
Die Anzahl der Operationen zum Bestimmen der Gewichtsmatrizen bei der zweiten Phase der Vergröberung unter Verwendung einer isometrischen richtungsabhängigen Clusterbasis $\{V_{tc}\}_{\substack{t \in \mathcal{T}_\mathcal{I} \\ c \in \mathcal{R}_t}}$ *ist beschränkt durch*

$$k^3 \mathcal{C}_{qr} \left((\mathcal{C}_{sk} + \mathcal{C}_{kk}) \left(\#\mathcal{T}_\mathcal{I} + p_\mathcal{I} \kappa^2 \mathcal{C}_w \right) + 1 \right) + k^2 \sum_{(t,s,c) \in \mathcal{M}_{\mathcal{I}\times\mathcal{I}}} \left(\mathcal{C}_{qr} (\#^{\mathfrak{I}} s) + 2(\#^{\mathfrak{I}} t) \right),$$

wobei $\mathcal{C}_w$ *durch*

$$\mathcal{C}_w := \max \left\{ \mathcal{C}_{kk} \mathcal{C}_{Ck} \mathcal{C}_{uk} \mathcal{C}_{mk}^2 |\Gamma|, \mathcal{C}_{tk} \right\}$$

gegeben ist.

Beweis: Sei ein $t \in \mathcal{T}_\mathcal{I}$ gegeben. Falls t an einem zulässigen Block (t, s, c) in Rang-k-Darstellung beteiligt ist, wird eine QR-Zerlegung von B_{ts} sowie eine anschließende Multiplikation mit der Dreiecksmatrix notwendig. Der Aufwand hierfür für alle Richtungen ist beschränkt durch

$$k^2 \sum_{c \in \mathcal{R}_t} \sum_{\substack{s \in \mathrm{row}_c(t) \\ (t,s,c) \in \mathcal{M}_{\mathcal{I}\times\mathcal{I}}}} \left(\mathcal{C}_{qr} (\#^{\mathfrak{I}} s) + 2(\#^{\mathfrak{I}} t) \right).$$

Das Weiterreichen zu den Kindern geschieht ausschließlich über Kopieren, entsprechend ist kein zusätzlicher Rechenaufwand nötig. Da die zweite Summe ohnehin schon auf Elemente $(t, s, c) \in \mathcal{M}_{\mathcal{I}\times\mathcal{I}}$ eingeschränkt ist, kann der Aufwand für die Gewichte der Rang-k-Darstellungen für alle Blöcke in $\mathcal{M}_{\mathcal{I}\times\mathcal{I}}$ durch

$$k^2 \sum_{(t,s,c) \in \mathcal{M}_{\mathcal{I}\times\mathcal{I}}} \left(\mathcal{C}_{qr} (\#^{\mathfrak{I}} s) + 2(\#^{\mathfrak{I}} t) \right)$$

procedure row_weights_coarsen($\mathcal{T}_{\mathcal{I}\times\mathcal{I}}, \mathcal{R}, t, V, S, Z, Rk, \widehat{A}, op$)
  Matrices T, Q, P, int m, n
  for all $c \in \mathcal{R}_t$ **do**
    $\widehat{A}_{tc} \leftarrow 0, \quad Z_{tc} \leftarrow 0$
    for $\tilde{s} \in \mathrm{row}_c^+(t)$ **do**
      if $op_{(t,\tilde{s},c)} ==$ true **then**
        Take Rk representation $(A_{t\tilde{s}}, B_{t\tilde{s}})$ and compute QR decomposition
        $B_{t\tilde{s}} = QT, \quad \mathbb{C}^{\mathfrak{I}t\times k_A} \ni \widehat{A}_{tc} \leftarrow (\widehat{A}_{tc} \quad A_{t\tilde{s}}T^*)$
      else
        $n \leftarrow m + k_{sc}, \qquad \mathbb{C}^{n\times k_{tc}} \ni T \leftarrow \begin{pmatrix} Z_{tc} \\ S_b^* \end{pmatrix}, \qquad m \leftarrow min\{n, k_{tc}\}$
        compute QR decomposition $T = QP, \quad \mathbb{C}^{m\times k_{tc}} \ni Z_{tc} \leftarrow P$
      end if
    end for
  end for
  if there is a parent cluster t^+ of t **then**
    for all $c \in \mathcal{R}_t$ **do**
      $\mathrm{vor}_t^{t^+}(c) = \{c_1^+, \ldots, c_\tau^+\}, \quad n \leftarrow m + \sum_{i=1}^{\tau} m_{(t^+)}(c_i^+)$ ▷ $m \widehat{=}$ rows Z_{tc}
      $\mathbb{C}^{n\times k_{tc}} \ni T \leftarrow \begin{pmatrix} Z_{t^+c_1^+}E^*_{tc_1^+} \\ \vdots \\ Z_{t^+c_\tau^+}E^*_{tc_\tau^+} \\ Z_{tc} \end{pmatrix}, \qquad m \leftarrow min\{n, k_{tc}\}$
      compute QR decomposition $T = QP, \qquad \mathbb{C}^{m\times k_{tc}} \ni Z_{tc} \leftarrow P$
      if $\widehat{A}_{t^+c}$ is not empty **then**
        For $\widehat{A}_{t^+c} \in \mathbb{C}^{\mathfrak{I}t^+\times k_A} \quad \widehat{A}_{tc} \leftarrow (\widehat{A}_{t^+c}|_{\star(t\times k_A)} \quad \widehat{A}_{tc})$
      end if
    end for
  end if
  for all $t' \in \mathrm{kind}(t)$ **do**
    row_weights_coarsen($\mathcal{T}_{\mathcal{I}\times\mathcal{I}}, \mathcal{R}, t', V, S, Z, Rk, \widehat{A}, op$)
  end for
end procedure

Algorithmus 6.4: Gewichtsmatrizen der Vergröberung

beschränkt werden. Die Anzahl der Operationen für die klassischen Gewichte der Clusterbasis kann grob durch den Aufwand der Gewichte aus der Rekompression (5.3.3) mit einer isometrischen Clusterbasis abgeschätzt werden, womit sich mit Bemerkung 32 direkt die Behauptung ergibt. □

In den Blättern angekommen, werden beide Komponenten der totalen richtungsabhängigen Clusterbasis für $Z_{tc} \in \mathbb{C}^{m \times k_{tc}}$ zusammengefügt

$$((V_{tc} Z_{tc}^*)|_{\star(t \times m)} \quad \widehat{A}_{tc}),$$

um dann analog zur Rekompression mit Hilfe von QR- und Singulärwertzerlegung den minimal nötigen Rang für die neue Clusterbasis zu ermitteln.
Anschließend werden die Basiswechselmatrizen bestimmt und für die Projektion der Kopplungsmatrizen gespeichert.
Hat der Elterncluster t^+ ebenfalls eine Matrix $\widehat{A}_{t^+c}$ für Anteile an Rang-k-Darstellungen, muss noch ein wenig mehr vorbereitet werden. Denn wenn der Cluster t^+ behandelt wird, kann $\widehat{A}_{t^+c}$ nicht einfach mit den in die neue Basis umgerechneten Transfermatrizen zusammengefügt werden, da $\widehat{A}_{t^+c}$ mit $\#\mathfrak{I}t^+$ Zeilen nicht zu den umgerechneten Transfermatrizen mit $r = \sum_{t \in \mathrm{kind}(t^+)} k_{tc}^{new}$ Zeilen passt. Entsprechend muss auch der Anteil der Rang-k-Darstellungen umgerechnet werden. Dies kann glücklicherweise direkt bei der Bearbeitung der Kinder als Vorbereitungsschritt durchgeführt werden. Der Teil von $\widehat{A}_{t^+c}$, der mit dem Kindercluster t und Richtung c in Verbindung steht, ist in den ersten Spalten von $\widehat{A}_{tc}$ gespeichert (siehe Algorithmus 6.4). Zum Umrechnen werden die auf den neuen Rang eingeschränkten linken Singulärvektoren $\widetilde{U} = Q|_{\star(r \times k_{tc}^{new})}$ aus der Singulärwertzerlegung von

$$((V_{tc} Z_{tc}^*)_{\star(t \times m)} \quad \widehat{A}_{tc}) = Q \Sigma O^*$$

verwendet. Wenn die neue Matrix der Clusterbasis für t bestimmt ist, kann $\widehat{A}_{tc}$ mit $\widetilde{U}^* \widehat{A}_{tc}$ überschrieben werden, so dass jetzt der umgerechnete Anteil in $\widehat{A}_{tc}$ zu finden ist. Wird der Elterncluster t^+ bearbeitet, kann die umgerechnete Variante A der Matrix $\widehat{A}_{t^+c}$ mit ihren k_A Spalten aus den schon umgerechneten Matrizen der Kinder $t_1, \ldots, t_\tau$ rekonstruiert werden

$$A = \begin{pmatrix} \widehat{A}_{t_1 c}|_{\star(k_{t_1 c}^{new} \times k_A)} \\ \vdots \\ \widehat{A}_{t_\tau c}|_{\star(k_{t_\tau c}^{new} \times k_A)} \end{pmatrix},$$

da die Anteile vom Elterncluster t^+ immer in den ersten Spalten der Matrizen $\widehat{A}_{tc}$ stehen. Die Singulärwertzerlegung wird dann von

$$(\widetilde{V} Z_{t^+c}^* \quad A),$$

bestimmt, wobei $\widetilde{V}$ aus den in die neue Basis umgerechneten Transfermatrizen besteht.
Falls weder der Elterncluster t^+ noch einer seiner Vorfahren an einer Rang-k-Matrix beteiligt sind, wird wie bei der Rekompression für alle Richtungen $c^+ \in \mathcal{R}_{t^+}$ eine Singulärwertzerlegung allein von $\widetilde{V} Z^*_{t^+c^+}$ bestimmt.

Da der Ablauf des Algorithmus zur Berechnung der neuen Clusterbasis nur minimal gegenüber der Rekompression verändert wurde, kann der Beweis der Komplexität leicht an den neuen Fall angepasst werden.
Sei S_v^{tc} die Teilmenge von S^{tc}, bei der nach der ersten Phase der Vergröberung die zugehörigen Matrizen als Rang-k-Matrix gespeichert werden. Für alle $s \in S_v^{tc}$ existiert also ein Vorfahre t^+ von t, so dass die zugehörige Teilmatrix $A|_{\star(t^+\times s)}$ über eine Rang-k-Darstellung $A_{t^+s}B^*_{t^+s}$ approximiert wird. Für die Abschätzung des Aufwands ist es wichtig, dass die Menge S_v^{tc} beschränkt ist. Sei $L_t \in \underline{p_{\mathcal{I}|}}$ die erste Stufe im Blockbaum, auf der ein zulässiger Vorfahre von t auftritt. Nehme an, dass für den entstehenden Baum $\mathcal{T}^{new}_{\mathcal{I}\times\mathcal{I}}$ eine Konstante $\mathcal{C}_{fk} \in \mathbb{N}_{\geq 1}$ existiert, so dass

$$\#S_v^{tc} \leq \mathcal{C}_{fk}(\operatorname{stufe}(t) - L_t + 1) \qquad \text{für alle } t \in \mathcal{T}_{\mathcal{I}}, c \in \mathcal{R}_t \tag{6.1.2}$$

erfüllt ist, also dass nur eine begrenzte Anzahl an Blöcken mit Verbindung zum Cluster t pro Stufe in Rang-k-Darstellung vorliegen. Dann kann die Anzahl der Spalten k_A der Gewichtsmatrizen für die Gewichte A der Rang-k-Matrizen mit

$$k_A \leq \#S_v^{tc} k \leq \mathcal{C}_{fk}(p_{\mathcal{I}} + 1)k$$

abgeschätzt werden.

Lemma 6.5 (Aufwand Berechnung Clusterbasis)
Die Anzahl der nötigen Operationen zum Bestimmen einer richtungsabhängigen Clusterbasis $\{V_{tc}\}_{\substack{t\in\mathcal{T}_{\mathcal{I}}\\ c\in\mathcal{R}_t}}$ *zur vergröberten* $\mathcal{RH}^2$*-Matrix, deren richtungsabhängiger Blockbaum* (6.1.2) *und* $p_{\mathcal{I}} \geq 1$ *erfüllt, ist durch*

$$k^3\mathcal{C}_{fk}(p_{\mathcal{I}}+1)\mathcal{C}_{vt}\left(\#\mathcal{T}_{\mathcal{I}} + \kappa^2(p_{\mathcal{I}}+1)\mathcal{C}_{tk}\right) \tag{6.1.3}$$

mit der Konstanten

$$\mathcal{C}_{vt} := \mathcal{C}_{kk} + 4\mathcal{C}_{sr} + \tfrac{3}{2}\mathcal{C}_{svd}\mathcal{C}_{sr}^2$$

beschränkt.

Beweis: Seien $t \in \mathcal{T}_{\mathcal{I}}$ und $c \in \mathcal{R}_t$ gegeben. Falls $\operatorname{kind}(t) = \emptyset$ erfüllt, wird direkt mit V_{tc} und A_{tc} gerechnet und es gilt $r = \#\mathfrak{I}t$.
Gilt hingegen $\operatorname{kind}(t) \neq \emptyset$, ist $r = \sum_{t'\in\operatorname{kind}(t)} k^{new}_{t'c'}$ und es muss für jedes Kind das Produkt

procedure trunc_coarsen($\mathcal{T}_\mathcal{I}, \mathcal{R}, t, V, V^{new}, Z, \widehat{A}, C, \epsilon, tm$)
Matrix $\widetilde{V}, A, X, Q, U, \Sigma, \widetilde{U}$, int k^{new}, r
for all $t' \in \mathrm{kind}(t)$ **do**
trunc_coarsen($\mathcal{T}_\mathcal{I}, \mathcal{R}, t', V, V^{new}, Z, \widehat{A}, C, \epsilon, tm$)
end for
for all directions $c \in \mathcal{R}_t$ **do**
$A \leftarrow 0$
if $\mathrm{kind}(t) = \emptyset$ **then**
$r \leftarrow \#^\mathfrak{I} t, \qquad \mathbb{C}^{r \times k_{tc}} \ni \widetilde{V} \leftarrow V_{tc}|_{\star(t \times k_{tc})}, \qquad \mathbb{C}^{r \times k_A} \ni A \leftarrow \widehat{A}_{tc}$
else $\mathrm{kind}(t) = \{t'_1, \ldots, t'_\tau\}$ with direction $c' = \mathrm{r}_t(c)$

$$r \leftarrow \sum_{i=1}^{\tau} k^{new}_{t'_i c'}, \qquad \mathbb{C}^{r \times k_{tc}} \ni \widetilde{V} \leftarrow \begin{pmatrix} C_{t'_1 c'} E_{t'_1 c} \\ \vdots \\ C_{t'_\tau c'} E_{t'_\tau c} \end{pmatrix}$$

If $\widehat{A}_{tc'} \neq 0$ with $\widehat{A}_{tc'} \in \mathbb{C}^{\mathfrak{I} t \times k_A}$ $\qquad \mathbb{C}^{r \times k_A} \ni A \leftarrow \begin{pmatrix} \widehat{A}_{t'_1 c'}|_{\star(k^{new}_{t'_1 c'} \times k_A)} \\ \vdots \\ \widehat{A}_{t'_\tau c'}|_{\star(k^{new}_{t'_\tau c'} \times k_A)} \end{pmatrix}$

end if
$\mathbb{C}^{r \times (\widetilde{m} + k_A)} \ni X \leftarrow (\widetilde{V} Z^*_{tc} \;\; A)$, compute SVD of X with $X = Q \Sigma O^*$ and
find minimal rank k^{new} for tm and ϵ $\qquad \triangleright \widetilde{m} \widehat{=}$ rows Z_{tc}
$\widetilde{U} \leftarrow Q|_{\star(r \times k^{new})}, \qquad \mathbb{C}^{k^{new} \times k_{tc}} \ni C_{tc} = \widetilde{U}^* \widetilde{V}, \qquad \widehat{A}_{tc} \leftarrow \widetilde{U}^* A$
if $\mathrm{kind}(t) = \emptyset$ **then**
$V^{new}_{tc}|_{\star(t \times k^{new})} \leftarrow \widetilde{U}$
else
$r = 0$ and $c' = \mathrm{r}_t(c)$
for all $t' \in \mathrm{kind}(t)$ **do**
$E^{new}_{t'c} \leftarrow \widetilde{U}|_{\star([r+1, r+k^{new}_{t'c'}] \times k^{new})}, \qquad r \leftarrow r + k^{new}_{t'c'}$
end for
end if
end for
end procedure

Algorithmus 6.5: Gemeinsame Clusterbasis bestimmen

der Basiswechselmatrix $C_{t'c'}$ und der Transfermatrix $E_{t'c}$ berechnet werden. Der Aufwand ist entsprechend durch

$$\sum_{t' \in \mathrm{kind}(t)} 2k_{t'c'}^{new} k_{t'c'} k_{tc} \leq 2k^3 \sum_{t' \in \mathrm{kind}(t)} 1 = 2k^3 \mathcal{C}_{kk}$$

beschränkt. Die Matrix A wird nur zusammenkopiert, so dass kein zusätzlicher Rechenaufwand entsteht. Für den ersten Block der Matrix X ist noch eine Multiplikation von V mit dem Gewicht nötig, was einen Aufwand von $2\widetilde{m} r k_{tc}$ hat. Die Singulärwertzerlegung hat einen Aufwand von $\mathcal{C}_{svd} r(\widetilde{m} + k_A) \min\{r, (\widetilde{m} + k_A)\}$. Die Bestimmung der Basiswechselmatrix ist durch $2k_{tc}^{new} r k_{tc}$ und die Vorbereitung der Matrix A_{tc} durch $2rk_{tc}^{new} k_A$ beschränkt. Abschätzen von $\min\{r, (\widetilde{m} + k_A)\}$ gegen r führt zu einem Aufwand von

$$2k^3 \mathcal{C}_{kk} + 2\widetilde{m} r k_{tc} + \mathcal{C}_{svd} r^2 (\widetilde{m} + k_A) + 2k_{tc}^{new} r k_{tc} + 2rk_{tc}^{new} k_A.$$

Unter Berücksichtigung von $\mathcal{C}_{sr} := \max\{\mathcal{C}_{bk}, \mathcal{C}_{kk}\}$ gilt $r \leq \mathcal{C}_{sr} k$ und mit $\widetilde{m} \leq k$ wird der Aufwand zu

$$\begin{aligned} &2k^3 \mathcal{C}_{kk} + 2\mathcal{C}_{sr} k^3 + \mathcal{C}_{svd} \mathcal{C}_{sr}^2 k^2 (k + k_A) + 2\mathcal{C}_{sr} k^3 + 2\mathcal{C}_{sr} k^2 k_A \\ &= 2k^3 (\mathcal{C}_{kk} + 2\mathcal{C}_{sr}) + \mathcal{C}_{svd} \mathcal{C}_{sr}^2 k^2 (k + k_A) + 2\mathcal{C}_{sr} k^2 k_A. \end{aligned}$$

Mit der Annahme 6.1.2 kann zusätzlich k_A abgeschätzt werden

$$\begin{aligned} &2k^3 (\mathcal{C}_{kk} + 2\mathcal{C}_{sr}) + \mathcal{C}_{svd} \mathcal{C}_{sr}^2 k^3 (1 + \mathcal{C}_{fk}(p_{\mathcal{I}} + 1)) + 2\mathcal{C}_{sr} k^3 \mathcal{C}_{fk}(p_{\mathcal{I}} + 1) \\ &= 2k^3 (\mathcal{C}_{kk} + 2\mathcal{C}_{sr} + \tfrac{1}{2} \mathcal{C}_{svd} \mathcal{C}_{sr}^2) + k^3 \mathcal{C}_{fk}(p_{\mathcal{I}} + 1)(2\mathcal{C}_{sr} + \mathcal{C}_{svd} \mathcal{C}_{sr}^2) \end{aligned}$$

und da $2 \leq \mathcal{C}_{fk}(p_{\mathcal{I}} + 1)$ gilt, kann dies durch

$$k^3 \mathcal{C}_{fk}(p_{\mathcal{I}} + 1) \underbrace{(\mathcal{C}_{kk} + 4\mathcal{C}_{sr} + \tfrac{3}{2} \mathcal{C}_{svd} \mathcal{C}_{sr}^2)}_{= \mathcal{C}_{vt}}$$

beschränkt werden. Für den Gesamtaufwand betrachte jedes $t \in \mathcal{T}_{\mathcal{I}}$ und jede Richtung $c \in \mathcal{R}_t^{eff}$, mit Lemma 3.14 ergibt sich folgendes Resultat

$$\begin{aligned} &\sum_{t \in \mathcal{T}_{\mathcal{I}}} \sum_{c \in \mathcal{R}_t^{eff}} k^3 \mathcal{C}_{fk}(p_{\mathcal{I}} + 1) \mathcal{C}_{vt} \\ &\leq k^3 \mathcal{C}_{fk}(p_{\mathcal{I}} + 1) \mathcal{C}_{vt} \left(\# \mathcal{T}_{\mathcal{I}} + \kappa^2 (p_{\mathcal{I}} + 1) \mathcal{C}_{tk} \right). \end{aligned}$$

□

Nachdem die neuen Clusterbasen bestimmt sind, werden im letzten Schritt die neuen Kopplungsmatrizen berechnet. Handelt es sich bei dem betrachteten Block b um einen noch im $\mathcal{RH}^2$-Matrix-Format vorliegenden Block, können die Basiswechselmatrizen C_{tc}, C_{sc} genutzt werden, um die alte Kopplungsmatrix zu projizieren

$$S_b^{new} = C_{tc} S_b C_{sc}^*.$$

Ist der Block in Rang-k-Darstellung gespeichert, werden mehr Multiplikationen nötig

$$S_b^{new} = \left((V_{tc}^{new}|_{\star(t\times k_{tc}^{new})})^* A_{ts} \right) \left(B_{ts}^* (W_{sc}^{new})|_{\star(s\times k_{sc}^{new})} \right).$$

Es ist vorteilhaft, zunächst je eine Matrix-Multiplikation der Rang-k-Darstellung mit der jeweiligen Matrix der Clusterbasis zu berechnen. Hierbei ist noch zu bedenken, dass die Matrizen $V_{tc}^{new}, W_{sc}^{new}$ nur in den Blättern direkt vorliegen und gegebenenfalls eine Vorwärtstransformation entlang der vererbten Richtung notwendig wird. Anschließend werden die beiden Zwischenprodukte multipliziert, um die neue Kopplungsmatrix zu erhalten.

---

procedure coarsen_project($\mathcal{T}_{\mathcal{I}\times\mathcal{I}}, \mathcal{R}, S, S^{new}, C_{row}, C_{col}, V^{new}, W^{new}, Rk, op$)
  Matrix T, P
  for $b = (t, s, c) \in \mathcal{L}^+_{\mathcal{I}\times\mathcal{I}}$ **do**
    if $op_b ==$ false **then**
      $\mathbb{C}^{k_{tc}^{new}\times k_{sc}^{new}} \ni S_b^{new} \leftarrow C_{tc} S_b C_{sc}^*$
    else
      Take corresponding Rk presentation with $Rk = (A_{ts}, B_{ts})$
      $T \leftarrow (V_{tc}^{new}|_{\star(t\times k_{tc}^{new})})^* A_{ts}$ if necesssary with forward transformation
      $P \leftarrow (W_{sc}^{new}|_{\star(s\times k_{sc}^{new})})^* B_{ts}$ if necesssary with forward transformation
      $\mathbb{C}^{k_{tc}^{new}\times k_{sc}^{new}} \ni S_b^{new} \leftarrow TP^*$
    end if
  end for
end procedure

Algorithmus 6.6: Bestimmen der neuen Kopplungsmatrizen

Lemma 6.6 (Aufwand Berechnung neue Kopplungsmatrizen beim Vergröbern)
Der Aufwand zur Berechnung der neuen Kopplungsmatrizen ist durch

$$2\,(2 + \mathcal{C}_{ecb})\, k^3 (\#\mathcal{I})^2$$

beschränkt.

Beweis: Betrachte einen zulässigen Block b. Falls für b noch eine alte Kopplungsmatrix gespeichert ist, sind zwei Multiplikationen nötig, der Aufwand kann damit durch $4k^3$ beschränkt werden.
Ist der Block hingegen in Rang-k-Darstellung abgelegt, werden mehr Multiplikationen nötig, mit Lemma 5.11 ist der Aufwand für die nötigen Vorwärtstransformationen entlang der vererbten Richtung c durch

$$k^2(\#^{\mathfrak{I}} t + \#^{\mathfrak{I}} s)\mathcal{C}_{ecb}$$

beschränkt. Die anschließende Multiplikation braucht nicht mehr als $2k^3$ Operationen. Da Cluster mindestens einen Index enthalten müssen, können beide Fälle durch

$$k^3(\#^{\mathfrak{I}}t + \#^{\mathfrak{I}}s)\,(2 + \mathcal{C}_{ecb})$$

beschränkt werden. Dies führt für alle Berechnungen von Kopplungsmatrizen zu

$$k^3\,(2 + \mathcal{C}_{ecb}) \sum_{(t,s,c)\in\mathcal{L}^{+,new}_{\mathcal{I}\times\mathcal{I}}} (\#^{\mathfrak{I}}t + \#^{\mathfrak{I}}s).$$

Mit Hilfe der Blattpartition 2.29 kann dies grob gegen

$$k^3\,(2 + \mathcal{C}_{ecb}) \sum_{(t,s,c)\in\mathcal{L}^{+,new}_{\mathcal{I}\times\mathcal{I}}} (\#^{\mathfrak{I}}t + \#^{\mathfrak{I}}s) \leq k^3\,(2 + \mathcal{C}_{ecb})\,2(\#\mathcal{I})^2$$

abgeschätzt werden. □

Der vollständige Vorgang des Vergröberns ist im Algorithmus 6.7 zusammengefasst.

---

procedure coarsen($\mathcal{T}_{\mathcal{I}\times\mathcal{I}}, \mathcal{R}, V, V^{new}, S, S^{new}, W, W^{new}, N, op, \epsilon_1, \epsilon_2, tm$)
  sets of matrices $Rk, \widehat{A}_{row}, Z_{row}, \widehat{A}_{col}, Z_{col}, C_{row}, C_{col}$
  build_coarsen ($\mathcal{T}_{\mathcal{I}\times\mathcal{I}}$, wurzel($\mathcal{T}_{\mathcal{I}\times\mathcal{I}}$), $V, S, W, N, Rk, \epsilon_1, op$) ▷ 6.3
  row_weights_coarsen ($\mathcal{T}_{\mathcal{I}\times\mathcal{I}}, \mathcal{R}$, wurzel($\mathcal{T}_{\mathcal{I}}$), $V, S, Z_{row}, Rk, \widehat{A}_{row}, op$) ▷ 6.4
  col_weights_coarsen ($\mathcal{T}_{\mathcal{I}\times\mathcal{I}}, \mathcal{R}$, wurzel($\mathcal{T}_{\mathcal{I}}$), $W, S, Z_{col}, Rk, \widehat{A}_{col}, op$)
  trunc_coarsen($\mathcal{T}_{\mathcal{I}}, \mathcal{R}$, wurzel($\mathcal{T}_{\mathcal{I}}$), $V, V^{new}, Z_{row}, \widehat{A}_{row}, C_{row}, \epsilon_2, tm$) ▷ 6.5
  trunc_coarsen ($\mathcal{T}_{\mathcal{I}}, \mathcal{R}$, wurzel($\mathcal{T}_{\mathcal{I}}$), $W, W^{new}, Z_{col}, \widehat{A}_{col}, C_{col}, \epsilon_2, tm$)
  coarsen_project ($\mathcal{T}_{\mathcal{I}\times\mathcal{I}}, \mathcal{R}, S, S^{new}, C_{row}, C_{col}, V^{new}, W^{new}, Rk, op$) ▷ 6.6
end procedure

Algorithmus 6.7: Vergröbern

Eine Schranke für den Aufwand des Vergröberns findet sich im nachstehenden Theorem.

Theorem 6.7 (Aufwand Vergröbern)
Der Gesamtaufwand des Vergröberns einer $\mathcal{RH}^2$-Matrix unter der Annahme, dass nur dann vergröbert wird, wenn $k \leq \#^{\mathfrak{I}}t, \#^{\mathfrak{I}}s$ für $t, s \in \mathcal{L}_{\mathcal{I}}$ gilt und der richtungsabhängige Blockbaum (6.1.2) *sowie $p_{\mathcal{I}} \geq 1$ erfüll, ist durch*

$$\begin{aligned} 2\,(2 + \mathcal{C}_{ecb})\,k^3(\#\mathcal{I})^2 + 4k^2(\#\mathcal{I})\mathcal{C}_{bk}\,(\mathcal{C}_{qr}(\mathcal{C}_{sk} + \mathcal{C}_{kk}) + \mathcal{C}_{fk}(p_{\mathcal{I}} + 1)\mathcal{C}_{vt}) \\ +2k^3(p_{\mathcal{I}} + 1)\kappa^2\,(\mathcal{C}_{qr}(\mathcal{C}_{sk} + \mathcal{C}_{kk})\mathcal{C}_w + \mathcal{C}_{fk}\mathcal{C}_{vt}(p_{\mathcal{I}} + 1)\mathcal{C}_{tk}) + 2k^3\mathcal{C}_{qr} \\ +k^2 \sum_{(t,s,c)\in\mathcal{M}_{\mathcal{I}\times\mathcal{I}}} (\mathcal{C}_{vv} + \mathcal{C}_{qr} + 2)\left(\#^{\mathfrak{I}}t + \#^{\mathfrak{I}}s\right) \end{aligned}$$

gegeben.

Beweis: Die Behauptung ergibt sich durch die Addition des Aufwands für die Berechnung von Rang-k-Darstellungen Gleichung (6.1.1), die Bestimmung der Gewichte für Zeilen- und Spaltenclusterbasis Lemma 6.4, die Berechnung der neuen Zeilen- und Spaltenclusterbasis Lemma 6.5 sowie die Berechnung der neuen Kopplungsmatrizen Lemma 6.6, wobei $\#\mathcal{T}_{\mathcal{I}}$ mit Korollar 3.17 abgeschätzt wird. □

Bemerkung 36 : *Beim Vergröbern von $\mathcal{H}^2$-Matrizen gibt es den Ansatz, statt mit Rang-k-Darstellungen als Intermediate mit Teil-$\mathcal{H}^2$-Matrizen zu arbeiten, die dann eigene Clusterbasen besitzen. Dadurch kann Speicher beim Vergröbern gespart werden. Leider lässt sich diese Idee aufgrund der Richtungen bei $\mathcal{RH}^2$-Matrizen nicht so leicht übertragen, da für einen Block mit zwei Kindern, die unterschiedliche Richtungen verwenden, eine Clusterbasis mit zwei Richtungen nötig wäre, was ungefähr dem Speicheraufwand von zwei Clusterbasen entspricht.*

6.2 Numerische Experimente

Bei den folgenden Experimenten geht es darum, zu zeigen, dass die erste Phase der Vergröberung eine angemessene Genauigkeit liefert sowie dass die zweite Phase zu einer drastischen Reduzierung des Speicherbedarfs führt. Insbesondere soll dabei die Speicherreduktion im Vergleich zur Rekompression gezeigt werden, die bis dahin zu den mit am speichereffizientesten Darstellungen geführt hat.
Alle Rechnungen wurden auf einem *Shared Memory* System mit zwei Intel® Xeon® Platinum 8160 Prozessoren mit insgesamt 48 Kernen durchgeführt. Bei allen Experimenten wurde mit $\eta_1 = 10$, $\eta_2 = 1$ und einer Auflösung von 32 gearbeitet, während die Interpolationsordnung variiert wurde.

Die Tabelle 6.1 zeigt Ergebnisse für den Fehler der ersten Phase der Vergröberung. Verglichen wurden hier die im Mischformat vorliegende vergröberte Matrix $\widetilde{A}_e^m$ mit der durch die Interpolation entstehenden $\widetilde{A}_e^i$. In der ersten Spalte ist die jeweilige Anzahl an Freiheitsgraden zu finden, gerechnet wurde mit stückweise konstanten Basispolynomen. Im oberen Teil der Tabelle wurde mit einer Interpolationsordnung von $m = 4$ im unteren Teil mit $m = 5$ gearbeitet (die beiden Teile der Tabelle sind durch eine Doppellinie getrennt).
Die zweite Spalte beinhalten die jeweils verwendete Wellenzahl κ, diese wurde so gewählt, dass das Problem als hochfrequent anzusehen ist ($h\kappa \approx 1.2$). Die verwendeten Genauigkeiten zum Vergröbern auf Mischformat sind in den Spalten drei und fünf. Der relative Fehler in der Frobeniusnorm zwischen dem Mischformat und der durch Interpolation entstehenden Approximation ist in Spalte vier und sechs zu finden.
Es zeigt sich, dass der Fehler nach der ersten Phase der Vergröberung mit der geforderten Genauigkeit fällt und damit, dass der Ansatz der vererbten Richtungen funktioniert.

n	κ	ϵ	$\frac{\|\tilde{A}_e^i - \tilde{A}_e^m\|_F}{\|\tilde{A}_e^i\|_F}$	ϵ	$\frac{\|\tilde{A}_e^i - \tilde{A}_e^m\|_F}{\|\tilde{A}_e^i\|_F}$
2048	8	1.0_{-3}	2.01_{-4}	1.0_{-4}	2.47_{-5}
2048	8	1.0_{-5}	2.28_{-6}	1.0_{-6}	2.48_{-7}
4608	12	1.0_{-3}	3.04_{-4}	1.0_{-4}	3.34_{-5}
4608	12	1.0_{-5}	3.47_{-6}	1.0_{-6}	3.42_{-7}
8192	16	1.0_{-3}	4.07_{-4}	1.0_{-4}	4.18_{-5}
8192	16	1.0_{-5}	4.27_{-6}	1.0_{-6}	4.50_{-7}
18432	24	1.0_{-3}	4.88_{-4}	1.0_{-4}	5.08_{-5}
18432	24	1.0_{-5}	5.13_{-6}	1.0_{-6}	5.23_{-7}

Tabelle 6.1: Fehler der ersten Phase der Vergröberung bei verschiedenen Größen n, Genauigkeiten ϵ und Wellenzahlen κ

Bei der zweiten Phase wird der Fehler in der relativen Frobeniusnorm im Vergleich zur vollbesetzten Matrix gemessen. Verglichen werden die Ergebnisse mit einer Approximation, die mit Hilfe der Rekompression entstanden ist und nicht mit vererbten Richtungen arbeitet. Dabei bezeichne mit A_e die vollbesetzte Matrix, mit $\widetilde{A}_e^v$ die vergröberte Approximation und mit $\widetilde{A}_e^r$ die rekomprimierte Approximation.

In der Tabellen 6.2, 6.3 und 6.4 befinden sich einige Ergebnisse für die Vergröberung bei verschiedenen Interpolationsordnungen und stückweise konstanten Basisfunktionen. In der ersten Spalte sind die Freiheitsgrade zu finden, in Spalte zwei die verwendete Genauigkeit für die Rekompression beziehungsweise für die zweite Phase der Vergröberung. Für die erste Phase der Vergröberung wurde durchgehend eine um den Faktor zehn kleinere Genauigkeit verwendet. In den Spalten vier und sieben ist der Speicherbedarf für die rekomprimierte beziehungsweise vergröberte Approximation zu finden. Die Spalten fünf und acht beinhalten die maximal auftretenden Ränge der beiden Approximationen. Der Fehler der Approximation gemessen in der relativen Frobeniusnorm in Bezug zur vollbesetzten Matrix ist in den Spalten sechs und neun zu finden.

Die Ergebnisse in den drei Tabellen zeigen, dass die Vergröberung ein ähnliches Konvergenzverhalten wie die normale Rekompression aufweist und ebenfalls den Rang gegenüber der Interpolation deutlich reduziert. Dabei ist der maximale Rang, der bei der Vergröberung auftritt, im Schnitt fast doppelt so groß wie bei der Rekompression, was dem erhöhten Informationsgehalt, den die Clusterbasis bei der Vergröberung tragen muss, geschuldet ist. Dafür ist der Gesamtspeicheraufwand der Vergröberung deutlich geringer und die Ersparnis wächst mit der steigenden Problemgröße.

n	ϵ	κ	$\widetilde{A}_e^r$ [KiB]/n	k_{max}	$\frac{\|A_e-\tilde{A}_e^r\|_F}{\|A_e\|_F}$	$\widetilde{A}_e^v$ [KiB]/n	k_{max}	$\frac{\|A_e-\tilde{A}_e^v\|_F}{\|A_e\|_F}$
2048	1.0_{-2}	8	32.4	3	1.04_{-4}	14.7	8	3.76_{-3}
2048	1.0_{-3}	8	32.4	5	1.52_{-5}	16.4	13	3.81_{-4}
2048	1.0_{-4}	8	32.5	8	1.00_{-6}	18.8	18	3.89_{-5}
2048	1.0_{-5}	8	32.5	9	2.68_{-7}	22.0	23	3.86_{-6}
4608	1.0_{-2}	12	72.4	5	1.76_{-4}	20.3	9	4.86_{-3}
4608	1.0_{-3}	12	71.7	7	1.53_{-5}	22.8	14	5.14_{-4}
4608	1.0_{-4}	12	71.8	10	3.99_{-6}	26.1	19	5.24_{-5}
4608	1.0_{-5}	12	71.9	13	3.21_{-6}	30.9	26	5.84_{-6}

Tabelle 6.2: Vergleich des Speicheraufwands und des Fehlers der Rekompression und Vergröberung für $m = 4$

n	ϵ	κ	$\widetilde{A}_e^r$ [KiB]/n	k_{max}	$\frac{\|A_e-\tilde{A}_e^r\|_F}{\|A_e\|_F}$	$\widetilde{A}_e^v$ [KiB]/n	k_{max}	$\frac{\|A_e-\tilde{A}_e^v\|_F}{\|A_e\|_F}$
8192	1.0_{-2}	16	119.5	6	5.04_{-4}	19.7	12	6.09_{-3}
8192	1.0_{-3}	16	120.3	9	5.48_{-5}	22.9	17	6.48_{-4}
8192	1.0_{-4}	16	121.3	12	5.11_{-6}	27.5	26	1.14_{-4}
8192	1.0_{-5}	16	122.9	17	1.25_{-6}	33.9	38	9.32_{-5}
18432	1.0_{-2}	12	103.7	6	2.65_{-3}	19.8	9	7.39_{-3}
18432	1.0_{-3}	12	111.4	10	3.24_{-4}	22.7	14	7.56_{-4}
18432	1.0_{-4}	12	125.1	14	3.34_{-5}	27.2	19	7.87_{-5}
18432	1.0_{-5}	12	144.2	19	7.63_{-6}	33.0	26	8.15_{-6}

Tabelle 6.3: Vergleich des Speicheraufwands und des Fehlers der Rekompression und Vergröberung für $m = 5$

n	ϵ	κ	$\widetilde{A}_e^r$ [KiB]/n	k_{max}	$\frac{\|A_e - \widetilde{A}_e^r\|_F}{\|A_e\|_F}$	$\widetilde{A}_e^v$ [KiB]/n	k_{max}	$\frac{\|A_e - \widetilde{A}_e^v\|_F}{\|A_e\|_F}$
32768	1.0_{-2}	32	286.6	6	1.86_{-3}	26.5	15	8.53_{-3}
32768	1.0_{-3}	32	296.0	10	2.10_{-4}	33.7	23	8.75_{-4}
32768	1.0_{-4}	32	310.5	14	2.12_{-5}	43.8	32	1.05_{-4}
32768	1.0_{-5}	32	332.2	18	2.06_{-6}	57.5	43	5.21_{-5}

Tabelle 6.4: Vergleich des Speicheraufwands und des Fehlers der Rekompression und Vergröberung für $m = 7$

Die Einsparungen im Speicher stammen aus der Reduktion von Nah- und Fernfeldmatrizen. Die Abbildung 6.1 zeigt dabei die Entwicklung des Speicherbedarfs der rekomprimierten Approximation im Vergleich zur Vergröberten.

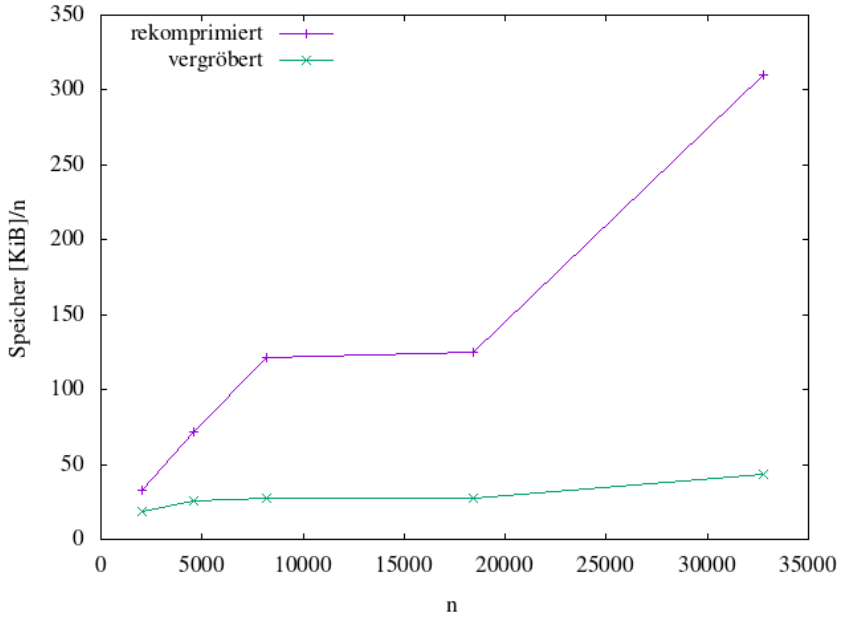


Abbildung 6.1: Vergleich des Speicherbedarfs pro Freiheitsgrad von Rekompression und Vergröberung

In Abbildung 6.1 ist noch einmal deutlich zu erkennen, dass der Speicherbedarf für die Vergröberung nahezu linear mit der Anzahl der Freiheitsgrade wächst. Ein Speicheraufwand in $\mathcal{O}(n)$ entspricht dem optimal Möglichen, die Vergröberung erfüllt damit die Erwartungen. Das Plateau in der Kurve der Rekompression rührt wohl daher, dass beim Experiment abhängig von der Anzahl der Freiheitsgrade unterschiedliche Interpolationsordnungen zum

Einsatz kamen. Diese sind dann teilweise am unteren Minimum, um die gewünschte Genauigkeit für die Rekompression zu gewährleisten, und teilweise weit darüber.
In der Tabelle 6.5 sind für die oben durchgeführten Experimente noch einmal die Anzahl der auftretenden Blöcke im richtungsabhängigen Blockbaum sowie die benötigte Zeit zum Vergröbern enthalten.

Wenn die Anzahl der genutzten Blöcke pro Freiheitsgrad verglichen wird, zeigt sich ein fast konstantes Verhalten der vergröberten Approximation im Gegensatz zum Standardfall. Durch die drastische Reduktion der vorkommenden Blöcke sind erhebliche Einsparungen im Speicher und damit einhergehend schnellere Matrixoperationen möglich.
Die Ergebnisse zeigen jedoch auch deutlich, dass die aktuelle Implementierung des Vergröberungsalgorithmus noch zu langsam ist. Ein großer Teil des Zeitaufwands ist auf die vielen nötigen Singulärwertzerlegungen zurückzuführen. Durch geschicktes Parallelisieren des Codes kann hier jedoch noch Zeit eingespart werden. Blöcke, die nicht in einem Vorfahren- oder Nachfahrenverhältnis zueinander stehen, können gut parallel zueinander bearbeitet werden, dabei ist nur Vorsicht bei den vielen Zugriffen und Veränderungen von gemeinsamen Strukturen geboten.
Der nächste Schritt ist entsprechend eine Überarbeitung der Implementierung hin zu höherer Parallelisierung.

n	κ	ϵ	$\widetilde{A}_e^r$ [s]	Blöcke	$\widetilde{A}_e^v$ [s]	Blöcke
2048	8	1.0_{-2}	0.8	11861	7.7	8277
2048	8	1.0_{-3}	0.7	11861	8.4	8277
2048	8	1.0_{-4}	0.7	11861	9.4	8277
2048	8	1.0_{-5}	0.7	11861	11.3	8277
4608	12	1.0_{-2}	1.9	39509	39.6	16989
4608	12	1.0_{-3}	1.8	39509	44.5	16989
4608	12	1.0_{-4}	1.7	39509	49.7	16989
4608	12	1.0_{-5}	1.8	39509	60.8	16989
8192	16	1.0_{-2}	9.6	129749	158.8	35381
8192	16	1.0_{-3}	9.8	129749	173.8	35381
8192	16	1.0_{-4}	10.1	129749	196.2	35381
8192	16	1.0_{-5}	10.3	129749	226.9	35389
18432	12	1.0_{-2}	122.3	465973	1437.3	84805
18432	12	1.0_{-3}	123.5	465973	1506.8	84805
18432	12	1.0_{-4}	124.6	465973	1554.6	84805
18432	12	1.0_{-5}	127.9	465973	1648.8	84805
32768	32	1.0_{-2}	1150.8	1705525	14106.4	155573
32768	32	1.0_{-3}	1185.5	1705525	14473.8	155573
32768	32	1.0_{-4}	1215.9	1705525	14784.3	155597
32768	32	1.0_{-5}	1245.2	1705525	15312.7	155909

Tabelle 6.5: Vergleich des Zeitaufwands der Rekompression und Vergröberung sowie der Anzahl der Blöcke

Ausblick für die Forschung

Mit dieser Arbeit wurde die Theorie und das Wissen um die $\mathcal{RH}^2$-Matrizen erweitert und ausgebaut. Die Schwachbesetztheit wurde mehr aus dem Blickwinkel der Kegel um die Richtungen betrachtet, eine erste Analyse des Interpolationsfehlers beim Doppelschichtoperator durchgeführt, Aussagen zum Fehler des Einfachschichtoperators auf Blöcke übertragen und Algorithmen für bessere Kompressionsraten und vergröberte Baumstrukturen vorgestellt, analysiert und im Rahmen der Programmbibliothek *H2Lib* programmiert sowie erste Tests realisiert.
Insgesamt sollte mit dieser Arbeit eine ausführliche Grundlage inklusive einer Zusammenfassung der bisherigen Forschungsergebnisse von Herrn Börm und Herrn Melenk auf dem Gebiet der $\mathcal{RH}^2$-Matrizen geschaffen worden sein. Das Lösen der Helmholtz-Gleichung über $\mathcal{RH}^2$-Matrizen bietet jedoch noch Raum für weitere und aufbauende Untersuchungen. Lohnenswerte Forschungsgebiete für die Zukunft wären eine Analyse des Hypersingulären Operators sowie eine tiefer gehende Analyse für den Doppelschichtoperator, um auch hier Fehleraussagen auf Blöcken treffen zu können. Auf der praktischen Seite ließen sich weiterführende Forschungen anschließen, die durch Parallelisieren und Modifizieren der Algorithmen zur Verbesserung der Kompressionsrate die Performanz steigern. Besonders der Vergröberungsalgorithmus mit seinem hohen Einsparpotential ist ein dankbarer Kandidat für aufbauende Forschungen.

Literaturverzeichnis

[1] BEATSON, R. ; GREENGARD, L. : A short course on fast multipole methods. In: *Wavelets, Multilevel Methods, and Elliptic PDEs*, Clarendon Press, 1997 (Numerical mathematics and scientific computation), S. 1–37

[2] BEBENDORF, M. ; KUSKE, C. ; VENN, R. : Wideband nested cross approximation for Helmholtz problems. In: *Numerische Mathematik* 130 (2015), Nr. 1, S. 1–34

[3] BÖRM, S. ; MELENK, J. M.: Approximation of the high-frequency Helmholtz kernel by nested directional interpolation: error analysis. In: *Numerische Mathematik* 137 (2017), Nr. 1, S. 1–34

[4] BÖRM, S. : *EMS Tracts in Mathematics*. Bd. 14: *Efficient Numerical Methods for Non-local Operators: $\mathcal{H}^2$-Matrix Compression, Algorithms and Analysis*. Zürich : EMS, 2010

[5] BÖRM, S. : *Wissenschaftliches Rechnen*. 2016. – https://www.informatik.uni-kiel.de/~sb/data/WissRech.pdf

[6] BÖRM, S. : Directional $\mathcal{H}^2$-matrix compression for high-frequency problems. In: *Numerical Linear Algebra with Application* 24 (2017), Nr. 6

[7] BÖRM, S. : *Numerik nicht-lokaler Operatoren*. 2018. – https://www.informatik.uni-kiel.de/~sb/data/NichtLokal.pdf

[8] BÖRM, S. : *Numerical Methods for Non-local Operators*. 2021. – https://www.informatik.uni-kiel.de/~sb/data/Nonlocal.pdf

[9] BÖRM, S. : *On iterated interpolation*. 2021. – https://arxiv.org/abs/2109.04330v1

[10] BÖRM, S. ; BÖRST, C. : Hybrid matrix compression for high-frequency problems. In: *SIAM Journal on Matrix Analysis and Applications* 41 (2020), Nr. 4, S. 1704–1725. http://dx.doi.org/10.1137/19M124280X. – DOI 10.1137/19M124280X

[11] BRAKHAGE, H. ; WERNER, P. : Über das Dirichletsche Außenraumproblem für die Helmholtzsche Schwingungsgleichung. In: *Archiv der Mathematik* 16 (1965), S. 325–329

[12] BRANDT, A. : Multilevel computations of integral transforms and particle interactions

with oscillatory kernels. In: *Comp. Phys. Comm.* 65 (1991), S. 24–38

[13] BURG, K. ; HAF, H. ; WILLE, F. ; MEISTER, A. : *Partielle Differentialgleichungen und funktionalanalytische Grundlagen*. 5., aktualisierte Auflage. Wiesbaden : Vieweg + Teubner Verlag, 2010

[14] BURGER, M. : *Mathematische Modellierungen - Wellen*. 2008. – Skript

[15] BURTON, A. J. ; MILLER, G. F.: The application of integral equation methods to the numerical solution of some exterior boundary-value problems. In: *Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences* 322 (1971), Nr. 1553, S. 201–210

[16] CARRIER, J. ; GREENGARD, L. ; ROKHLIN, V. : A fast adaptive multipole algorithm for particle simulations. In: *SIAM Journal on Scientific and Statistical Computing* 9 (1988), Nr. 4, S. 669–686

[17] CHANDLER-WILDE, S. N. ; GRAHAM, I. G. ; LANGDON, S. ; SPENCE, E. A.: Numerical-asymptotic boundary integral methods in high-frequency acoustic scattering. In: *Acta Numerica* 21 (2012), S. 89–305

[18] DAHMEN, W. ; REUSKEN, A. : *Numerik für Ingenieure und Naturwissenschaftler*. Zweite, korrigierte Auflage. Berlin Heidelberg : Springer-Verlag, 2008

[19] DEUFLHARD, P. ; HOHMANN, A. : *Numerische Mathematik 1, Eine algorithmisch orientierte Einführung*. 4. Auflage. Berlin : De Gruyter, 2008

[20] DEVORE, R. A. ; LORENTZ, G. G.: *Constructive Approximation*. Berlin : Springer-Verlag, 1993 (Grundlehren der mathematischen Wissenschaften)

[21] DOBROWOLSKI, M. : *Angewandte Funktionalanalysis: Funktionalanalysis, Sobolev-Räume und elliptische Differentialgleichungen*. 2., korrigierte und überarbeitete Auflage. Berlin Heidelberg : Springer Verlag, 2010

[22] ENGLEDER, S. : *Stabilisierte Randintegralgleichungen für äussere Randwertprobleme der Helmholtz-Gleichung*, Technische Universität Graz, Diplomarbeit, 2006

[23] ENGQUIST, B. ; YING, L. : Fast directional multilevel algorithms for oscillatory kernels. In: *SIAM Journal on Scientific Computing* 29 (2007), Nr. 4, S. 1710–1737

[24] FREUND, R. W. ; HOPPE, R. H.: *Stoer/Bulirsch: Numerische Mathematik 1*. Zehnte, neu bearbeitete Auflage. Berlin Heidelberg : Springer-Verlag, 2007

[25] GERLACH, W. : Helmholtz, Hermann von. In: *Hess - Hüttig*. Verlag Duncker & Humblot, 1969 (Neue Deutsche Biographie 8), S. 498–501

[26] GOLUB, G. H. ; VAN LOAN, C. F.: *Matrix Computations*. 3. Edition. London : The

Johns Hopkins University Press, 1996

[27] GRASDYCK, L. : *Theorie und Anwendung hierarchischer Matrizen*, Christian-Albrechts-Universität zu Kiel, Diss., 2001

[28] GREENGARD, L. ; ROKHLIN, V. : A fast algorithm for particle simulations. In: *Journal of Computational Physics* 73 (1987), Nr. 2, S. 325–348

[29] HACKBUSCH, W. : A Sparse Matrix Arithmetic Based on H-Matrices. Part I: Introduction to H-Matrices. In: *Computing* 62 (1999), Nr. 2, S. 89–108

[30] HACKBUSCH, W. ; GRASEDYCK, L. : Construction and Arithmetics of H-Matrices. In: *Computing* 70 (2003), Nr. 4, S. 295–334

[31] HACKBUSCH, W. ; NOWAK, Z. : On the fast matrix multiplication in the boundary element method by panel clustering. In: *Numerische Mathematik* 54 (1989), Nr. 4, S. 463–491

[32] HACKBUSCH, W. : *Hierarchische Matrizen - Algorithmen und Analysis*. Berlin, Heidelberg : Springer-Verlag, 2009

[33] HACKBUSCH, W. : *Theorie und Numerik elliptischer Differentialgleichungen*. 4. überarbeitete Auflage. Wiesbaden : Springer Spektrum, 2017

[34] HANKE-BOURGEOIS, M. : *Grundlagen der Numerischen Mathematik und des Wissenschaftlichen Rechnens*. 3. aktualisierte Auflage. Wiesbaden : Vieweg + Teubner Verlag, 2009

[35] HOUSEHOLDER, A. S.: Unitary Triangularization of a Nonsymmetric Matrix. In: *Journal of the ACM* 5 (1958), Nr. 4, S. 339–342

[36] JÄNICH, K. : *Funktionentheorie*. 6. Auflage. Berlin, Heidelberg : Springer, 2004

[37] MASON, J. C. ; HANDSCOMB, D. C.: *Chebyshev Polynomials*. Boca Raton : CRC Press, 2003

[38] MCLEAN, W. : *Strongly Elliptic Systems and Boundary Integral Equations*. New York : Cambridge University Press, 2000

[39] MESSNER, M. ; SCHANZ, M. ; DARVE, E. : Fast directional multilevel summation for oscillatory kernels based on Chebyshev interpolation. In: *Journal of Computational Physics* 231 (2012), Nr. 4, S. 1175–1196

[40] MICHIELSSEN, E. ; BOAG, A. : A multilevel matrix decomposition algorithm for analyzing scattering from large structures. In: *IEEE Transactions on Antennas and Propagation* 44 (1996), Nr. 8, S. 1086–1093

[41] NEHARI, Z. : *Conformal Mapping*. University of California : McGraw-Hill Book Company, 1952. – reprinted in 1975 by Dover Publications, Inc. New York

[42] RELLICH, F. : Über das asymptotische Verhalten der Lösungen von $\Delta u + \lambda u = 0$ in unendlichen Gebieten. In: *Jahresbericht der Deutschen Mathematiker-Vereinigung* 53 (1943), S. 57–65

[43] RIVLIN, T. J.: *Chebyshev polynomials : From Approximation Theory to Algebra and Number Theory*. 2nd. New York : Wiley, 1990

[44] ROKHLIN, V. : Rapid solution of integral equations of classical potential theory. In: *Journal of Computational Physics* 60 (1985), Nr. 2, S. 187–207

[45] ROKHLIN, V. : Diagonal Forms of Translation Operators for the Helmholtz Equation in Three Dimensions. In: *Applied and Computational Harmonic Analysis* 1 (1993), Nr. 1, S. 82–93

[46] SAYAS, F. : *Introduction to the boundary element method. A case study: the Helmholtz equation*. classnotes, 2006. – https://team-pancho.github.io/documents/escuelaChile.pdf

[47] SOMMERFELD, A. : Die Greensche Funktion der Schwingungsgleichung. In: *Jahresbericht der Deutschen Mathematiker-Vereinigung* 21 (1912), S. 309–352

[48] STEINBACH, O. : Boundary integral equations for Helmholtz boundary value and transmission problems. In: *Berichte aus dem Institut für Numerische Mathematik*, Technische Universität Graz, 2012

[49] TRÖLTZSCH, F. : *Optimale Steuerung partieller Differentialgleichungen - Theorie, Verfahren und Anwendungen*. 1. Wiesbaden : Vieweg+Teubner Verlag, 2005

Abbildungsverzeichnis

Algorithmenverzeichnis

Symbolverzeichnis

$\mathcal{H}$	Bezeichnung für eine hierarchische Matrix, die Teilmatrizen auf (t, s) durch ein Niedrigrang-Produkt $A_{ts}B_{ts}^*$ approximiert, siehe Einführung von Kapitel 2 und Definition 6.1
$\mathcal{H}^2$	Bezeichnung für eine hierarchische Matrix mit zweifacher Hierarchie, die zulässige Teilmatrizen auf (t, s) über ein Produkt $V_t S_{t,s} W_s^*$ approximiert, wobei V_t, W_s jeweils Elemente einer geschachtelten Clusterbasis sind, siehe Einführung von Kapitel 2
$\mathcal{RH}^2$	Bezeichnung für eine richtungsabhängige Variante der $\mathcal{H}^2$-Matrix, siehe Definition 3.4
$Id \in \mathbb{C}^{n\times n}$	Bezeichnung für die Einheitsmatrix im $\mathbb{C}^{n\times n}$ für $n \in \mathbb{N}$
$\mathrm{diag}(\sigma_1, \ldots, \sigma_n) \in \mathbb{R}^{n\times n}$	Diagonalmatrix im $\mathbb{R}^{n\times n}$ mit Diagonaleinträgen $\sigma_1, \ldots, \sigma_n$.
$\mathrm{i} \in \mathbb{C}$	Bezeichnung für die komplexe Einheit, $\mathrm{i} := \sqrt{-1}$
$\underline{n}$	Kurzform für die Menge der natürlichen Zahle von 1 bis n, also $\underline{n} := \{1, \ldots, n\}$
$\underline{n}_m$ für $m \in \mathbb{N}_0, m \leq n$	Erweiterung der Menge $\underline{n}$ auf einen ganzzahligen nicht negativen Startwert $m \leq n$, entspricht damit der Menge der natürlichen Zahle von m bis n, also $\underline{n}_m := \{m, \ldots, n\}$
$\mathcal{R}$	Familie der Richtungsmengen, siehe Definition 2.19
$\mathcal{R}_{\mathrm{stufe}\,(t)}$	Menge aller Richtungen auf der zu dem Cluster t gehörenden Stufe, siehe Definition 2.19
$\mathcal{I}$	Eine Indexmenge, siehe Kapitel 2.4
$\mathcal{T}_{\mathcal{I}}$	Ein Clusterbaum zu der Indexmenge $\mathcal{I}$, siehe Definition 2.13 2.20
$\mathcal{L}_{\mathcal{I}}$	Blätter eines Clusterbaums zu der Indexmenge $\mathcal{I}$, siehe Definition 2.16
$\mathcal{T}_{\mathcal{I}\times\mathcal{I}}$	Ein Blockbaum zu der Indexmenge $\mathcal{I} \times \mathcal{I}$, siehe Definition 2.25

$\mathcal{L}_{\mathcal{I}\times\mathcal{I}}$	Blätter eines Blockbaums zur Indexmenge $\mathcal{I} \times \mathcal{I}$, siehe Definition 2.26
$\mathcal{L}^-_{\mathcal{I}\times\mathcal{I}}$	Zulässige Blöcke der Menge $\mathcal{L}_{\mathcal{I}\times\mathcal{I}}$, siehe Definition 2.31
$\mathcal{L}^+_{\mathcal{I}\times\mathcal{I}}$	Unzulässige Blöcke der Menge $\mathcal{L}_{\mathcal{I}\times\mathcal{I}}$, siehe Definition 2.31
wurzel	Wurzel eines Baums, siehe Einführung des Kapitels 2.4
dist	Funktion zum Messen des Abstands zwischen zwei Quadern, siehe Kapitel 2.3
diam	Funktion zum Messen des Durchmessers eines Quaders, siehe Kapitel 2.3
kind	Menge der Kinder eines Clusters/Blocks, siehe Einführung zu Kapitel 2.4
nac	Menge der Nachfahren eines Clusters, siehe Definition 2.17
$\mathrm{supp}(f)$	Träger der Funktion f
stufe	Stufe eines Clusters/Blocks im Baum, siehe Definition 2.14 2.26
row	Spaltenclusterpartner, Menge aller Spaltencluster, die mit diesem Zeilencluster zusammen einen Block bilden, siehe Kapitel 2.5
row_c	Richtungsabhängige Variante von row, also die Einschränkung von row auf die Richtung c, siehe Definition 2.30
col	Zeilenclusterpartner, Menge aller Zeilencluster, die mit diesem Spaltencluster zusammen einen Block bilden, siehe Kapitel 2.5
col_c	Richtungsabhängige Variante von col, also Einschränkung von col auf die Richtung c, siehe Definition 2.30
vor	Menge der Vorfahren eines Clusters, siehe Definition 2.17
r_c	Bestapproximation der Richtung c in der nächst höheren Stufe, siehe Definition 2.19
$\mathfrak{I}t$	Menge der korrespondierenden Indizes zu einem Cluster t, siehe Einführung zu Kapitel 2.4
$A\|_{t\times s}$	Mit Nullen aufgefüllte Einschränkung der Matrix A auf die zu (t, s) korrespondierenden Indizes, siehe Definition 3.1
$A\|_{\star(t\times s)}$	Einschränkung der Matrix A auf die zu (t, s) korrespondierenden Indizes, siehe Definition 3.1

Erklärung

Hiermit erkläre ich, dass ich die vorliegende Dissertation - abgesehen von der Beratung durch meinen Betreuer Herrn Prof. Dr. Steffen Börm - nach Inhalt und Form eigenständig und nur mit den angegebenen Hilfsmitteln verfasste habe. Dabei habe ich die Regeln guter wissenschaftlicher Praxis der *Deutschen Forschungsgemeinschaft* eingehalten.
Die Arbeit hat weder ganz noch in Teilen an einer anderen Stelle im Rahmen eines Prüfungsverfahrens vorgelegen. Die Veröffentlichung, auf denen Teile der Dissertation basieren, ist auf Seite vii aufgeführt und die entsprechenden Stellen in der Arbeit sind gekennzeichnet.

Ich versichere außerdem, dass diesem Promotionsverfahren keine endgültig gescheiterten Promotionsverfahren vorausgegangen sind und dass keine akademischen Grade entzogen wurden.

Kiel, den 14.12.2021